

JFE

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re U.S. Patent Application of

MUROTANI et al.

Application Number: 10/821,888

Filed: April 12, 2004

For: DISK ARRAY APPARATUS AND DISK ARRAY
APPARATUS CONTROL METHOD

ATTORNEY DOCKET NO. ASAM.0119



Honorable Assistant Commissioner
for Patents
Washington, D.C. 20231

LETTER

Sir:

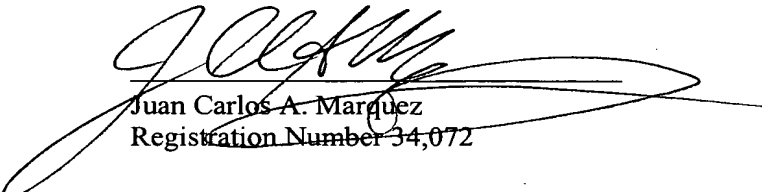
The below-identified communications are submitted in the above-captioned application or proceeding:

- | | | |
|---|--|---|
| <input checked="" type="checkbox"/> (X) | Priority Document (1) | |
| <input type="checkbox"/> () | Request for Priority | <input type="checkbox"/> () Assignment Document |
| <input type="checkbox"/> () | Response to Missing Parts
w/ signed Declaration | <input type="checkbox"/> () Petition under 37 C.F.R. 1.47(a) |
| | | <input type="checkbox"/> () Check for \$ |

- ☒ [X] The Commissioner is hereby authorized to charge payment of any fees associated with this communication, including fees under 37 C.F.R. § 1.16 and 1.17 or credit any overpayment to Deposit Account Number 08-1480. A duplicate copy of this sheet is attached.

Respectfully submitted,

Stanley P. Fisher
Registration Number 24,344



Juan Carlos A. Marquez
Registration Number 34,072

REED SMITH LLP
3110 Fairview Park Drive
Suite 1400
Falls Church, Virginia 22042
(703) 641-4200
July 7, 2004

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re U.S. Patent Application of

MUROTANI et al.

Application Number: 10/821,888

Filed: April 12, 2004

For: DISK ARRAY APPARATUS AND DISK ARRAY
APPARATUS CONTROL METHOD

ATTORNEY DOCKET NO. ASAM.0119



Honorable Assistant Commissioner
for Patents
Washington, D.C. 20231

**REQUEST FOR PRIORITY
UNDER 35 U.S.C. § 119
AND THE INTERNATIONAL CONVENTION**

Sir:

In the matter of the above-captioned application for a United States patent, notice is hereby given that the Applicant claims the priority date of February 16, 2004, the filing date of the corresponding Japanese patent application 2004-038169.

A certified copy of Japanese patent application 2004-038169, is being submitted herewith. Acknowledgment of receipt of the certified copy is respectfully requested in due course.

Respectfully submitted,

Stanley P. Fisher
Registration Number 24,344


Juan Carlos A. Marquez
Registration Number 34,072

REED SMITH LLP
3110 Fairview Park Drive
Suite 1400
Falls Church, Virginia 22042
(703) 641-4200
July 7, 2004

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 4 年 2 月 1 6 日
Date of Application:

出 願 番 号 特 願 2 0 0 4 - 0 3 8 1 6 9
Application Number:
[ST. 10/C]: [J P 2 0 0 4 - 0 3 8 1 6 9]

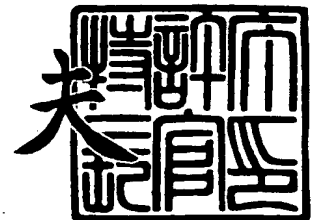
出 願 人 株式会社日立製作所
Applicant(s):



2 0 0 4 年 4 月 1 4 日

特許庁長官
Commissioner,
Japan Patent Office

今 井 康 夫



出証番号 出証特 2 0 0 4 - 3 0 3 1 3 2 1

【書類名】 特許願
【整理番号】 340301446
【提出日】 平成16年 2月16日
【あて先】 特許庁長官殿
【国際特許分類】 G06F 3/06
【発明者】
 【住所又は居所】 神奈川県小田原市中里 3 2 2 番 2 号 株式会社日立製作所 R A
 I D システム事業部内
 【氏名】 室谷 暁
【発明者】
 【住所又は居所】 神奈川県小田原市中里 3 2 2 番 2 号 株式会社日立製作所 R A
 I D システム事業部内
 【氏名】 石川 篤
【発明者】
 【住所又は居所】 神奈川県小田原市中里 3 2 2 番 2 号 株式会社日立製作所 R A
 I D システム事業部内
 【氏名】 岸本 哲哉
【特許出願人】
 【識別番号】 000005108
 【氏名又は名称】 株式会社日立製作所
【代理人】
 【識別番号】 110000176
 【氏名又は名称】 一色国際特許業務法人
 【代表者】 一色 健輔
【手数料の表示】
 【予納台帳番号】 211868
 【納付金額】 21,000円
【提出物件の目録】
 【物件名】 特許請求の範囲 1
 【物件名】 明細書 1
 【物件名】 図面 1
 【物件名】 要約書 1

【書類名】 特許請求の範囲**【請求項 1】**

情報処理装置と通信可能に接続され、
一又は複数の論理ボリュームが形成されている第一の記憶デバイスと、
一又は複数の論理ボリュームが形成されている第二の記憶デバイスと、
第三の記憶デバイスと
に対するデータの書き込みまたは読み出しを行うディスク制御装置であって、
メモリと、
計時機構と
を有し、

前記第一の記憶デバイスの前記論理ボリュームの識別子を主論理ボリュームの識別子、
前記第二の記憶デバイスの前記論理ボリュームの識別子を副論理ボリュームの識別子とする
対応付けが前記メモリにペア管理テーブルとして記憶され、

前記情報処理装置から前記第一の記憶デバイスの前記論理ボリュームに対するデータの
書き込み要求と前記データとを受信する書き込み要求受信部と、

前記書き込み要求を受信すると前記データを前記第一の記憶デバイスの前記論理ボリ
ュームに書き込む第一の書き込み部と、

前記データが書き込まれている前記第一の記憶デバイスの前記論理ボリュームの識別子
、当該論理ボリューム内の前記データが格納されている位置情報、前記計時機構より取得
される現在時刻である更新時刻、及び前記データで構成されるジャーナルデータを前記第
三の記憶デバイスに書き込むジャーナル書き込み部と、

前記第三の記憶デバイスに記憶されている前記ジャーナルデータの前記更新時刻を参照
し、前記計時機構より取得される現在時刻と前記更新時刻との差が所定の時間以上である
前記ジャーナルデータを選択し、選択された前記ジャーナルデータの前記更新時刻が早い
順に前記ジャーナルデータの前記論理ボリュームの識別子と前記位置情報と前記データと
を参照し、当該論理ボリュームの識別子が前記主論理ボリュームの識別子である前記副論
理ボリュームの識別子を前記ペア管理テーブルから取得し、当該データを前記第二の記憶
デバイスの当該副論理ボリュームの識別子で示される前記論理ボリュームの当該位置情報
で示される場所に書き込む第二の書き込み部と
を有することを特徴とするディスク制御装置。

【請求項 2】

請求項 1 に記載のディスク制御装置において、

前記第一の記憶デバイスを構成する一又は複数のハードディスクドライブと、前記第二
の記憶デバイス及び前記第三の記憶デバイスを構成する一又は複数のハードディスクドラ
イブとが物理的に別であることを特徴とするディスク制御装置。

【請求項 3】

請求項 1 に記載のディスク制御装置において、

前記所定の時間が前記メモリに記憶され、

前記第二の書き込み部は、前記メモリに記憶されている前記所定の時間を参照し、前記
ジャーナルデータの前記更新時刻と前記計時機構より取得される現在時刻との差が前記所
定の時間以上である前記ジャーナルデータを選択すること
を特徴とするディスク制御装置。

【請求項 4】

請求項 1 に記載のディスク制御装置において、

前記第一の記憶デバイスを構成する一又は複数のハードディスクドライブと、前記第二
の記憶デバイスを構成する一又は複数のハードディスクドライブとが物理的に別であるこ
とを特徴とするディスク制御装置。

【請求項 5】

請求項 1 に記載のディスク制御装置において、

前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを中断するス

プリット指示命令を受信するスプリット指示命令受信部と、

前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを再開するスプリット解除命令を受信するスプリット解除命令受信部と、

前記スプリット指示命令を受信すると前記スプリット指示命令を受信したことを示すデータと前記計時機構より取得される現在時刻であるスプリット時刻とで構成されるスプリット指示データを前記第三の記憶デバイスに書き込むスプリット指示記憶部と、

前記スプリット解除命令を受信すると前記スプリット解除命令を受信したことを示すデータであるスプリット解除データを前記第三の記憶デバイスに書き込むスプリット解除記憶部と

を有し、

前記第二の書き込み部は前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、前記スプリット解除データが記憶されていない場合は、前記スプリット指示データの前記スプリット時刻を参照し、前記更新時刻が当該スプリット時刻より遅い前記ジャーナルデータの前記データを前記第二の記憶デバイスに書き込まないことを特徴とするディスク制御装置。

【請求項 6】

請求項 5 に記載のディスク制御装置において、

前記第二の記憶デバイスの前記論理ボリュームに対応する仮想的な論理ボリュームである仮想論理ボリュームが設けられ、前記論理ボリュームの識別子と前記仮想論理ボリュームの識別子との対応付けが前記メモリに仮想論理ボリューム管理テーブルとして記憶され

、
前記第三の記憶デバイスを参照し、前記スプリット指示データが記憶され、かつ、前記スプリット解除データが記憶されていない場合は、前記更新時刻が前記スプリット指示データの前記スプリット時刻より早い前記ジャーナルデータの前記論理ボリュームの識別子と前記位置情報とで構成される未反映情報を前記メモリに記憶する未反映情報記憶部と、

前記情報処理装置から前記仮想論理ボリュームの識別子と位置情報とが設定されているデータの読み出し要求を受信する読み出し要求受信部と、

前記読み出し要求の前記仮想論理ボリュームの識別子に対応する前記第二の記憶デバイスの前記論理ボリュームの識別子を前記仮想論理ボリューム管理テーブルから取得する識別子取得部と、

前記メモリに記憶されている前記未反映情報を参照し、前記識別子取得部により取得された前記第二の記憶デバイスの前記論理ボリュームの識別子と前記未反映情報の前記論理ボリュームの識別子とが同じ前記未反映情報において、前記読み出し要求に設定されている前記位置情報が示す範囲と前記未反映情報の前記位置情報が示す範囲とを比較して重複する範囲を取得する重複範囲取得部と、

前記読み出し要求に設定されている前記位置情報が示す範囲のうち前記重複する範囲については前記第三の記憶デバイスに記憶されている前記ジャーナルデータの前記データを読み出し、前記読み出し要求に設定されている前記位置情報が示す範囲のうち前記重複する範囲以外の範囲については前記識別子取得部により取得された前記第二の記憶デバイスの前記論理ボリュームの識別子で示される前記論理ボリュームに記憶されている前記データを読み出す仮想論理ボリューム読み出し部と、

前記仮想論理ボリューム読み出し部により読み出された前記データを前記情報処理装置に送信する読み出しデータ送信部と
を有することを特徴とするディスク制御装置。

【請求項 7】

請求項 5 に記載のディスク制御装置において、

前記第二の書き込み部は前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、前記スプリット解除データが記憶されていない場合は、前記更新時刻が当該スプリット時刻より早い前記ジャーナルデータの前記データを、前記更新時刻と前記計時機構より取得される現在時刻との差が前記所定の時間以上であるかどうかにかかわらず、前

記第二の記憶デバイスに書き込むこと
を特徴とするディスク制御装置。

【請求項 8】

請求項 1 に記載のディスク制御装置において、

前記第一の記憶デバイスの前記論理ボリュームにグループ ID が付与され、前記論理ボリュームの識別子と前記グループ ID との対応付けが前記メモリにグループ ID 管理テーブルとして記憶され、

前記ジャーナル書き込み部は、前記第一の記憶デバイスの前記データが書き込まれている前記論理ボリュームの識別子に対応する前記グループ ID を前記グループ ID 管理テーブルから取得し、当該グループ ID を前記ジャーナルデータに設定して前記第三の記憶デバイスに書き込み、

前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを中断するスプリット指示命令を受信するスプリット指示命令受信部と、

前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを再開するスプリット解除命令を受信するスプリット解除命令受信部と、

前記スプリット指示命令を受信すると前記スプリット指示命令に設定されている前記グループ ID と前記スプリット指示命令を受信したことを示すデータと前記計時機構より取得される現在時刻であるスプリット時刻とで構成されるスプリット指示データを前記第三の記憶デバイスに書き込むスプリット指示記憶部と、

前記スプリット解除命令を受信すると前記スプリット解除命令に設定されている前記グループ ID と前記スプリット解除命令を受信したことを示すデータであるスプリット解除データを前記第三の記憶デバイスに書き込むスプリット解除記憶部と

を有し、

前記第二の書き込み部は、前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、当該スプリット指示データの前記グループ ID と同じ前記グループ ID が設定されている前記スプリット解除データが記憶されていない場合は、前記スプリット指示データの前記グループ ID と前記スプリット時刻とを参照し、前記スプリット指示データの前記グループ ID が 前記ジャーナルデータの前記グループ ID が前記スプリット指示データの当該グループ ID と同じで、かつ、前記更新時刻が当該スプリット時刻より遅い前記ジャーナルデータの前記データを前記第二の記憶デバイスに書き込まないことを特徴とするディスク制御装置。

【請求項 9】

請求項 8 に記載のディスク制御装置において、

前記第二の記憶デバイスの前記論理ボリュームに対応する仮想的な論理ボリュームである仮想論理ボリュームが設けられ、前記論理ボリュームの識別子と前記仮想論理ボリュームの識別子との対応付けが前記メモリに仮想論理ボリューム管理テーブルとして記憶され、

前記第三の記憶デバイスを参照し、前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、当該スプリット指示データの前記グループ ID と同じ前記グループ ID が設定されている前記スプリット解除データが記憶されていない場合は、当該スプリット指示データの前記グループ ID と同じ前記グループ ID が設定され、かつ、前記更新時刻が前記スプリット指示データの前記スプリット時刻より早い前記ジャーナルデータの前記論理ボリュームの識別子と前記位置情報とで構成される未反映情報を前記メモリに記憶する未反映情報記憶部と、

前記情報処理装置から前記仮想論理ボリュームの識別子と位置情報とが設定されているデータの読み出し要求を受信する読み出し要求受信部と、

前記読み出し要求の前記仮想論理ボリュームの識別子に対応する前記第二の記憶デバイスの前記論理ボリュームの識別子を前記仮想論理ボリューム管理テーブルから取得する識別子取得部と、

前記メモリに記憶されている前記未反映情報を参照し、前記識別子取得部により取得さ

れた前記第二の記憶デバイスの前記論理ボリュームの識別子と前記未反映情報の前記論理ボリュームの識別子とが同じ前記未反映情報において、前記読み出し要求に設定されている前記位置情報が示す範囲と前記未反映情報の前記位置情報が示す範囲とを比較して重複する範囲を取得する重複範囲取得部と、

前記読み出し要求に設定されている前記位置情報が示す範囲のうち前記重複する範囲については前記第三の記憶デバイスに記憶されている前記ジャーナルデータに設定されている前記データを読み出し、前記読み出し要求に設定されている前記位置情報が示す範囲のうち前記重複する範囲以外の範囲については前記識別子取得部により取得された前記第二の記憶デバイスの前記論理ボリュームの識別子で示される前記論理ボリュームに記憶されている前記データを読み出す仮想論理ボリューム読み出し部と、

前記仮想論理ボリューム読み出し部により読み出された前記データを前記情報処理装置に送信する読み出しデータ送信部と
を有することを特徴とするディスク制御装置。

【請求項 10】

請求項 8 に記載のディスク制御装置において、

前記第二の書き込み部は、前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、当該スプリット指示データの前記グループ ID と同じ前記グループ ID が設定されている前記スプリット解除データが記憶されていない場合は、前記ジャーナルデータの前記グループ ID が前記スプリット指示データの当該グループ ID と同じで、かつ、前記更新時刻が当該スプリット時刻より早い前記ジャーナルデータの前記データを、前記更新時刻と前記計時機構より取得される現在時刻との差が前記所定の時間以上であるかどうかにかかわらず、前記第二の記憶デバイスに書き込むこと
を特徴とするディスク制御装置。

【請求項 11】

情報処理装置と通信可能に接続され、

一又は複数の論理ボリュームが形成されている第一の記憶デバイスと、

一又は複数の論理ボリュームが形成されている第二の記憶デバイスと、

第三の記憶デバイスと

に対するデータの書き込みまたは読み出しを行うディスク制御装置の制御方法であって

メモリと、

計時機構と

を有し、

前記第一の記憶デバイスの前記論理ボリュームの識別子を主論理ボリュームの識別子、前記第二の記憶デバイスの前記論理ボリュームの識別子を副論理ボリュームの識別子とする対応付けが前記メモリにペア管理テーブルとして記憶され、

前記情報処理装置から前記第一の記憶デバイスの前記論理ボリュームに対するデータの書き込み要求と前記データとを受信するステップと、

前記書き込み要求を受信すると前記データを前記第一の記憶デバイスの前記論理ボリュームに書き込むステップと、

前記データが書き込まれている前記第一の記憶デバイスの前記論理ボリュームの識別子、当該論理ボリューム内の前記データが格納されている位置情報、前記計時機構より取得される現在時刻である更新時刻、及び前記データで構成されるジャーナルデータを前記第三の記憶デバイスに書き込むステップと、

前記第三の記憶デバイスに記憶されている前記ジャーナルデータの前記更新時刻を参照し、前記計時機構より取得される現在時刻と前記更新時刻との差が所定の時間以上である前記ジャーナルデータを選択し、選択された前記ジャーナルデータの前記更新時刻が早い順に前記ジャーナルデータの前記論理ボリュームの識別子と前記位置情報と前記データとを参照し、当該論理ボリュームの識別子が前記主論理ボリュームの識別子である前記副論理ボリュームの識別子を前記ペア管理テーブルから取得し、当該データを前記第二の記憶

デバイスの当該論理ボリュームの識別子で示される前記論理ボリュームの当該位置情報で示される場所に書き込むステップと

を有することを特徴とするディスク制御装置の制御方法。

【請求項 12】

請求項 11 に記載のディスク制御装置の制御方法において、

前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを中断するスプリット指示命令を受信するステップと、

前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを再開するスプリット解除命令を受信するステップと、

前記スプリット指示命令を受信すると前記スプリット指示命令を受信したことを示すデータと前記計時機構より取得される現在時刻であるスプリット時刻とで構成されるスプリット指示データを前記第三の記憶デバイスに書き込むステップと、

前記スプリット解除命令を受信すると前記スプリット解除命令を受信したことを示すデータであるスプリット解除データを前記第三の記憶デバイスに書き込むステップと

を有し、

前記第三の記憶デバイスに記憶されている更新データを前記第二の記憶デバイスの前記論理ボリュームに書き込む前記ステップは、前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、前記スプリット解除データが記憶されていない場合は、前記スプリット指示データの前記スプリット時刻を参照し、前記更新時刻が当該スプリット時刻より遅い前記ジャーナルデータの前記データを前記第二の記憶デバイスに書き込まないステップであること

を特徴とするディスク制御装置の制御方法。

【請求項 13】

請求項 12 に記載のディスク制御装置の制御方法において、

前記第二の記憶デバイスの前記論理ボリュームに対応する仮想的な論理ボリュームである仮想論理ボリュームが設けられ、前記論理ボリュームの識別子と前記仮想論理ボリュームの識別子との対応付けが前記メモリに仮想論理ボリューム管理テーブルとして記憶され、

前記第三の記憶デバイスを参照し、前記スプリット指示データが記憶され、かつ、前記スプリット解除データが記憶されていない場合は、前記更新時刻が前記スプリット指示データの前記スプリット時刻より早い前記ジャーナルデータの前記論理ボリュームの識別子と前記位置情報とで構成される未反映情報を前記メモリに記憶するステップと、

前記情報処理装置から前記仮想論理ボリュームの識別子と位置情報とが設定されているデータの読み出し要求を受信するステップと、

前記読み出し要求の前記仮想論理ボリュームの識別子に対応する前記第二の記憶デバイスの前記論理ボリュームの識別子を前記仮想論理ボリューム管理テーブルから取得するステップと、

前記メモリに記憶されている前記未反映情報を参照し、前記識別子取得部により取得された前記第二の記憶デバイスの前記論理ボリュームの識別子と前記未反映情報の前記論理ボリュームの識別子とが同じ前記未反映情報において、前記読み出し要求に設定されている前記位置情報が示す範囲と前記未反映情報の前記位置情報が示す範囲とを比較して重複する範囲を取得するステップと、

前記読み出し要求に設定されている前記位置情報が示す範囲のうち前記重複する範囲については前記第三の記憶デバイスに記憶されている前記ジャーナルデータの前記データを読み出し、前記読み出し要求に設定されている前記位置情報が示す範囲のうち前記重複する範囲以外の範囲については前記識別子取得部により取得された前記第二の記憶デバイスの前記論理ボリュームの識別子で示される前記論理ボリュームに記憶されている前記データを読み出すステップと、

前記仮想論理ボリューム読み出し部により読み出された前記データを前記情報処理装置に送信するステップと

を有することを特徴とするディスク制御装置の制御方法。

【請求項 14】

請求項 12 に記載のディスク制御装置の制御方法において、

前記第三の記憶デバイスに記憶されている更新データを前記第二の記憶デバイスの前記論理ボリュームに書き込む前記ステップは、前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、前記スプリット解除データが記憶されていない場合は、前記更新時刻が当該スプリット時刻より早い前記ジャーナルデータの前記データを、前記更新時刻と前記計時機構より取得される現在時刻との差が前記所定の時間以上であるかどうかにかかわらず、前記第二の記憶デバイスに書き込むステップであることを特徴とするディスク制御装置の制御方法。

【請求項 15】

請求項 11 に記載のディスク制御装置の制御方法において、

前記第一の記憶デバイスの前記論理ボリュームにグループ ID が付与され、前記論理ボリュームの識別子と前記グループ ID との対応付けが前記メモリにグループ ID 管理テーブルとして記憶され、

前記ジャーナルデータを前記第三の記憶デバイスに書き込む前記ステップは、前記第一の記憶デバイスの前記データが書き込まれている前記論理ボリュームの識別子に対応する前記グループ ID を前記グループ ID 管理テーブルから取得し、当該グループ ID を前記ジャーナルデータに設定して前記第三の記憶デバイスに書き込むステップであり、

前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを中断するスプリット指示命令を受信するステップと、

前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを再開するスプリット解除命令を受信するステップと、

前記スプリット指示命令を受信すると前記スプリット指示命令に設定されている前記グループ ID と前記スプリット指示命令を受信したことを示すデータと前記計時機構より取得される現在時刻であるスプリット時刻とで構成されるスプリット指示データを前記第三の記憶デバイスに書き込むステップと、

前記スプリット解除命令を受信すると前記スプリット解除命令に設定されている前記グループ ID と前記スプリット解除命令を受信したことを示すデータであるスプリット解除データを前記第三の記憶デバイスに書き込むステップと

を有し、

前記第三の記憶デバイスに記憶されている更新データを前記第二の記憶デバイスの前記論理ボリュームに書き込む前記ステップは、前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、当該スプリット指示データの前記グループ ID と同じ前記グループ ID が設定されている前記スプリット解除データが記憶されていない場合は、前記スプリット指示データの前記グループ ID と前記スプリット時刻とを参照し、前記スプリット指示データの前記グループ ID が 前記ジャーナルデータの前記グループ ID が前記スプリット指示データの当該グループ ID と同じで、かつ、前記更新時刻が当該スプリット時刻より遅い前記ジャーナルデータの前記データを前記第二の記憶デバイスに書き込まないステップであることを

を特徴とするディスク制御装置の制御方法。

【請求項 16】

請求項 15 に記載のディスク制御装置の制御方法において、

前記第二の記憶デバイスの前記論理ボリュームに対応する仮想的な論理ボリュームである仮想論理ボリュームが設けられ、前記論理ボリュームの識別子と前記仮想論理ボリュームの識別子との対応付けが前記メモリに仮想論理ボリューム管理テーブルとして記憶され、

前記第三の記憶デバイスを参照し、前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、当該スプリット指示データの前記グループ ID と同じ前記グループ ID が設定されている前記スプリット解除データが記憶されていない場合は、当該スプリ

ット指示データの前記グループIDと同じ前記グループIDが設定され、かつ、前記更新時刻が前記スプリット指示データの前記スプリット時刻より早い前記ジャーナルデータの前記論理ボリュームの識別子と前記位置情報とで構成される未反映情報を前記メモリに記憶するステップと、

前記情報処理装置から前記仮想論理ボリュームの識別子と位置情報とが設定されているデータの読み出し要求を受信するステップと、

前記読み出し要求の前記仮想論理ボリュームの識別子に対応する前記第二の記憶デバイスの前記論理ボリュームの識別子を前記仮想論理ボリューム管理テーブルから取得するステップと、

前記メモリに記憶されている前記未反映情報を参照し、前記識別子取得部により取得された前記第二の記憶デバイスの前記論理ボリュームの識別子と前記未反映情報の前記論理ボリュームの識別子とが同じ前記未反映情報において、前記読み出し要求に設定されている前記位置情報が示す範囲と前記未反映情報の前記位置情報が示す範囲とを比較して重複する範囲を取得するステップと、

前記読み出し要求に設定されている前記位置情報が示す範囲のうち前記重複する範囲については前記第三の記憶デバイスに記憶されている前記ジャーナルデータに設定されている前記データを読み出し、前記読み出し要求に設定されている前記位置情報が示す範囲のうち前記重複する範囲以外の範囲については前記識別子取得部により取得された前記第二の記憶デバイスの前記論理ボリュームの識別子で示される前記論理ボリュームに記憶されている前記データを読み出すステップと、

前記仮想論理ボリューム読み出し部により読み出された前記データを前記情報処理装置に送信するステップと

を有することを特徴とするディスク制御装置の制御方法。

【請求項 17】

請求項 15 に記載のディスク制御装置の制御方法において、

前記第三の記憶デバイスに記憶されている更新データを前記第二の記憶デバイスの前記論理ボリュームに書き込む前記ステップは、前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、当該スプリット指示データの前記グループIDと同じ前記グループIDが設定されている前記スプリット解除データが記憶されていない場合は、前記ジャーナルデータの前記グループIDが前記スプリット指示データの当該グループIDと同じで、かつ、前記更新時刻が当該スプリット時刻より早い前記ジャーナルデータの前記データを、前記更新時刻と前記計時機構より取得される現在時刻との差が前記所定の時間以上であるかどうかにかかわらず、前記第二の記憶デバイスに書き込むステップであること

を特徴とするディスク制御装置の制御方法。

【請求項 18】

情報処理装置と通信可能に接続され、

一又は複数のハードディスクドライブにより構成される第一の記憶デバイスと、

一又は複数のハードディスクドライブにより構成される第二の記憶デバイスと、

一又は複数のハードディスクドライブにより構成される第三の記憶デバイスと

に対するデータの書き込みまたは読み出しを行うディスク制御装置であって、

前記第一の記憶デバイスに一又は複数の論理ボリュームが形成され、

前記第二の記憶デバイスに一又は複数の論理ボリュームが形成され、

前記情報処理装置から前記論理ボリュームに対するデータの書き込み要求を受信するチャンネル制御部と、

前記チャンネル制御部により受信された前記データを前記論理ボリュームに書き込むディスク制御部と、

共有メモリと、

前記チャンネル制御部と前記ディスク制御部と前記共有メモリとを通信可能に接続するスイッチング制御部と

を有し、

前記第一の記憶デバイスの前記論理ボリュームの識別子を主論理ボリュームの識別子、前記第二の記憶デバイスの前記論理ボリュームの識別子を副論理ボリュームの識別子とする対応付けが前記メモリにペア管理テーブルとして記憶され、

前記ディスク制御部は、

前記チャネル制御部により受信された前記書き込み要求に伴い前記第一の記憶デバイスの前記論理ボリュームに前記データを書き込むと、前記データが書き込まれている前記第一の記憶デバイスの前記論理ボリュームの識別子、当該論理ボリューム内の前記データが格納されている位置情報、前記計時機構より取得される現在時刻である更新時刻、及び前記データで構成されるジャーナルデータを前記第三の記憶デバイスに書き込むジャーナル書き込み部と、

前記第三の記憶デバイスに記憶されている前記ジャーナルデータの前記更新時刻を参照し、前記計時機構より取得される現在時刻と前記更新時刻との差が前記共有メモリに記憶されている検出時間以上である前記ジャーナルデータを選択し、選択された前記ジャーナルデータの前記更新時刻が早い順に前記ジャーナルデータの前記論理ボリュームの識別子と前記位置情報と前記データとを参照し、当該論理ボリュームの識別子が前記主論理ボリュームの識別子である前記副論理ボリュームの識別子を前記ペア管理テーブルから取得し、当該データを前記第二の記憶デバイスの当該副論理ボリュームの識別子で示される前記論理ボリュームの当該位置情報で示される場所へ書き込む第二の書き込み部とを有することを特徴とするディスク制御装置。

【書類名】明細書**【発明の名称】** ディスクアレイ装置及びディスクアレイ装置の制御方法**【技術分野】****【0001】**

本発明は、ディスクアレイ装置及びディスクアレイ装置の制御方法に関する。

【背景技術】**【0002】**

近年、インターネットを利用した商品販売システムなど、24時間無停止で運用される情報処理システムが増加している。このような情報処理システムで用いられるデータは、データ量の増加や耐障害性の向上の要求に伴い、ディスクアレイ装置上のデータベースに格納されることが多い。無停止で運用される情報処理システムにおいては、データベースの更新処理を停止することなく、データベースに格納されているデータをバックアップする必要がある。そこで、ディスクアレイ装置内にバックアップ用のレプリカボリュームを設け、データベースに格納されているデータをバックアップ用のレプリカボリュームにも記憶しておく方法が用いられている。レプリカボリュームへのデータの書き込みを停止し、レプリカボリュームに記憶されているデータを磁気テープ等の外部記憶媒体にコピーすることにより、データベースのある時点のバックアップを取得することができる。

【0003】

特許文献1においては、レプリカボリュームへのデータの書き込みを停止している間にデータベースに書き込まれたデータを、バックアップの完了後にレプリカボリュームに書き込むリシンク処理が開示されている。

リシンク処理を実行している最中は、レプリカボリュームの内容は保証されない。そこで、1つのデータベースに対して2つのレプリカボリュームを設け、リシンク処理を2つのレプリカボリュームで交互に行っている場合もある。この場合、データベースを格納するデータ領域にハードウェア障害が発生すると、いずれか一方のレプリカボリュームとデータベースのREDOログとを用いてデータベースを復旧することが可能である。

【特許文献1】 米国特許第6141773号明細書**【発明の開示】****【発明が解決しようとする課題】****【0004】**

データベースの運用においては、上記のハードウェア障害の他に、ソフトウェアの不具合やユーザの操作ミス等によりデータベースに不正なデータが書き込まれる場合がある。このような障害をソフトウェア障害と呼んでいる。ソフトウェア障害が発生すると、磁気テープ等に保存されているある時点のデータを復元し、復元されたデータとREDOログとを用いてデータベースを復旧する必要がある。磁気テープ等からのデータ復元には相当の時間を要するため、システムの停止時間が長くなってしまう。

【0005】

そこで、ソフトウェア障害が発生した際に、磁気テープ等の外部記憶媒体からデータを復元することなく、ディスクアレイ装置内のデータを用いて迅速にデータベースを復旧することが求められている。ここで、前述した2つのレプリカボリュームを用いる場合、データベースを格納するボリュームの3倍の記憶容量が必要となり、ディスクアレイ装置を導入する際の費用が増大する。

【0006】

そのため、ハードウェア障害及びソフトウェア障害に対して、データベースをバックアップするために必要となる記憶容量を低減させることが求められている。

また、リシンク処理においては、レプリカボリュームに書き込むためのデータをデータベースから読み出す必要がある。そのため、リシンク処理中は外部からデータベースへのアクセス性能が劣化するという問題がある。そこで、外部からデータベースへのアクセス性能を低下させずに、データベースを迅速に復旧することが求められている。

また、データベースの迅速な復旧とは別に、データベースの更新処理を停止することな

く、万が一に備えてデータベースのある時点のバックアップを取得することが求められている。

【0007】

本発明は上記課題を鑑みてなされたものであり、ディスクアレイ装置及びディスクアレイ装置の制御方法を提供することを主たる目的とする。

【課題を解決するための手段】

【0008】

上記課題を解決するために、本発明のうち主たる発明では、情報処理装置と通信可能に接続され、一又は複数の論理ボリュームが形成されている第一の記憶デバイスと、一又は複数の論理ボリュームが形成されている第二の記憶デバイスと、第三の記憶デバイスとに対するデータの書き込みまたは読み出しを行うディスク制御装置であって、メモリと、計時機構とを有し、前記第一の記憶デバイスの前記論理ボリュームの識別子を主論理ボリュームの識別子、前記第二の記憶デバイスの前記論理ボリュームの識別子を副論理ボリュームの識別子とする対応付けが前記メモリにペア管理テーブルとして記憶され、前記情報処理装置から前記第一の記憶デバイスの前記論理ボリュームに対するデータの書き込み要求と前記データとを受信する書き込み要求受信部と、前記書き込み要求を受信すると前記データを前記第一の記憶デバイスの前記論理ボリュームに書き込む第一の書き込み部と、前記データが書き込まれている前記第一の記憶デバイスの前記論理ボリュームの識別子、当該論理ボリューム内の前記データが格納されている位置情報、前記計時機構より取得される現在時刻である更新時刻、及び前記データで構成されるジャーナルデータを前記第三の記憶デバイスに書き込むジャーナル書き込み部と、前記第三の記憶デバイスに記憶されている前記ジャーナルデータの前記更新時刻を参照し、前記計時機構より取得される現在時刻と前記更新時刻との差が所定の時間以上である前記ジャーナルデータを選択し、選択された前記ジャーナルデータの前記更新時刻が早い順に前記ジャーナルデータの前記論理ボリュームの識別子と前記位置情報と前記データとを参照し、当該論理ボリュームの識別子が前記主論理ボリュームの識別子である前記副論理ボリュームの識別子を前記ペア管理テーブルから取得し、当該データを前記第二の記憶デバイスの当該副論理ボリュームの識別子で示される前記論理ボリュームの当該位置情報で示される場所へ書き込む第二の書き込み部とを有することとする。

【発明の効果】

【0009】

ディスクアレイ装置及びディスクアレイ装置の制御方法を提供することができる。

【発明を実施するための最良の形態】

【0010】

==ディスクアレイ装置==

本実施の形態におけるディスクアレイ装置の第一の形態を図1に示す。ディスクアレイ装置10は、ディスク制御装置110と複数のハードディスクドライブ120とを備えている。ディスクアレイ装置10は情報処理装置20と通信手段により接続されている。通信手段は、例えば、LAN (Local Area Network) や SAN (Storage Area Network)、iSCSI (Internet Small Computer System Interface)、ESCON (Enterprise Systems Connection) (登録商標)、FICON (Fibre Connection) (登録商標) などである。

【0011】

情報処理装置20は、CPU (Central Processing Unit) やメモリを備えるコンピュータであり、パーソナルコンピュータやワークステーション、メインフレームなどのコンピュータである。情報処理装置20は、結合された複数台のコンピュータで構成されることもある。情報処理装置20ではオペレーティングシステムが動作している。オペレーティングシステム上ではアプリケーションソフトウェアが動作している。アプリケーションソフトウェアは、例えば、銀行の自動預金預け払いシステムや航空機の座席予約システムの機能を提供する。

【0012】

ディスク制御装置 110 はディスクアレイ装置 10 全体の制御を司る。ディスク制御装置 110 は、情報処理装置 20 から受信したコマンドに従ってハードディスクドライブ 120 に対する制御を行う。例えば情報処理装置 20 からデータの入出力要求を受信して、ハードディスクドライブ 120 に記憶されているデータの入出力のための処理を行う。

【0013】

ディスク制御装置 110 は、チャンネル制御部 131、ディスク制御部 132、共有メモリ 133、キャッシュメモリ 134、これらの間を通信可能に接続するクロスバスイッチなどで構成されるスイッチング制御部 135、及び管理端末 136などを備えて構成される。また、ディスク制御装置 110 を構成する各部 131～136 は耐障害性を高めるために冗長化されてもよい。

【0014】

キャッシュメモリ 134 は、主としてチャンネル制御部 131 とディスク制御部 132 との間で授受されるデータを一時的に記憶するために用いられる。例えばチャンネル制御部 131 が情報処理装置 20 から受信したデータ入出力コマンドが書き込みコマンドである場合には、チャンネル制御部 131 は情報処理装置 20 から受信した書き込みデータをキャッシュメモリ 134 に書き込む。またディスク制御部 132 はキャッシュメモリ 134 から書き込みデータを読み出してハードディスクドライブ 120 に書き込む。なお、キャッシュメモリ 134 は不揮発とすることもできる。この場合、チャンネル制御部 131 が情報処理装置 20 から受信したデータがキャッシュメモリ 134 に書き込まれた段階で、情報処理装置 20 に書込完了通知を送信することとしてもよい。

【0015】

ディスク制御部 132 は、チャンネル制御部 131 により共有メモリ 133 に書き込まれたデータ入出力要求を読み出してそのデータ入出力要求に設定されているコマンド（例えば、SCSI (Small Computer System Interface) 規格のコマンド）に従ってハードディスクドライブ 120 にデータの書き込みや読み出しなどの処理を実行する。ディスク制御部 132 はハードディスクドライブ 120 から読み出したデータをキャッシュメモリ 134 に書き込む。またデータの書き込み完了通知や読み出し完了通知などをチャンネル制御部 131 に送信する。ディスク制御部 132 は、ハードディスクドライブ 120 をいわゆる RAID (Redundant Array of Inexpensive Disks) 方式に規定される RAID レベル（例えば、0, 1, 5）で制御する機能を備えることもある。

【0016】

ハードディスクドライブ 120 により提供される記憶領域は、この記憶領域上に論理的に設定されるボリュームである論理ボリューム 121 を単位として管理されている。ハードディスクドライブ 120 へのデータの書き込みや読み出しは、論理ボリュームに付与される識別子を指定して行なうことができる。

【0017】

管理端末 136 はディスクアレイ装置 10 を保守・管理するためのコンピュータである。チャンネル制御部 131 やディスク制御部 132 において実行されるソフトウェアやパラメータの変更は、管理端末 136 からの指示により行われる。管理端末 136 はディスクアレイ装置 10 に内蔵される形態とすることもできるし、別体とすることもできる。

【0018】

共有メモリ 133 はチャンネル制御部 131、ディスク制御部 132、及び管理端末 136 からアクセスが可能である。チャンネル制御部 131 とディスク制御部 132 との間におけるデータ入出力要求コマンドの受け渡しに利用される他、ディスクアレイ装置 10 の管理情報等が記憶される。

【0019】

図 2 はチャンネル制御部 131 の構成を示すブロック図である。チャンネル制御部 131 は、インタフェース部 201、メモリ 202、CPU 203、NVRAM (Nonvolatile Random-Access Memory) 204、コネクタ 205 を備え、これらが一枚もしくは複数枚の回

路基板に一体的なユニットとして形成されている。

【0020】

インタフェース部201は、情報処理装置20との間で通信を行うためのインタフェースを備えている。通信を行うためのインタフェースとは、例えば、ファイバチャネルに対応したコネクタやイーサネット（登録商標）に対応したコネクタなどである。

【0021】

コネクタ205は、チャンネル制御部131をディスクアレイ装置10に接続するためのコネクタである。コネクタ205がディスクアレイ装置10側のコネクタと嵌合することにより、チャンネル制御部131が形成される基板はディスクアレイ装置10と電氣的に接続される。チャンネル制御部131はコネクタ205を介してスイッチング制御部135に接続され、共有メモリ133、キャッシュメモリ134、ディスク制御部132などにアクセスすることができる。

【0022】

CPU203はチャンネル制御部131の全体の制御を司る。CPU203はメモリ202やNVRAM203に格納されている各種プログラムを実行することによりチャンネル制御部131の機能を実現する。NVRAM203は各種プログラムや設定データなどを格納する不揮発性のメモリである。NVRAM203に記憶される各種プログラムや設定データなどの内容は管理端末136からの指示により書き換えを行うことができる。

【0023】

図3はディスク制御部132の構成を示すブロック図である。ディスク制御部132は、インタフェース部301、メモリ302、CPU303、NVRAM304、コネクタ305を備え、これらが一枚もしくは複数枚の回路基板に一体的なユニットとして形成されている。

【0024】

インタフェース部301は、ハードディスクドライブ120との間で通信を行うためのインタフェースを備えている。通信を行うためのインタフェースとは、例えば、SCSI規約に従うコネクタやファイバチャネル規約に従うコネクタなどである。

【0025】

コネクタ305は、ディスク制御部132をディスクアレイ装置10に接続するためのコネクタである。コネクタ305がディスクアレイ装置10側のコネクタと嵌合することにより、ディスク制御部132が形成される基板はディスクアレイ装置10と電氣的に接続される。ディスク制御部132はコネクタ305を介してスイッチング制御部135に接続され、共有メモリ133、キャッシュメモリ134、チャンネル制御部131などにアクセスすることができる。

【0026】

CPU303はディスク制御部132の全体の制御を司る。CPU303はメモリ302やNVRAM303に格納されている各種プログラムを実行することによりディスク制御部132の機能を実現する。NVRAM303は各種プログラムや設定データなどを格納する不揮発性のメモリである。NVRAM303に記憶される各種プログラムや設定データなどの内容は管理端末136からの指示により書き換えを行うことができる。

【0027】

次に、本実施の形態におけるディスクアレイ装置の第二の形態を図4に示す。第二の形態においては、ディスク制御装置110の構成が図1に示した第一の形態と異なる。ディスク制御装置110は、CPU141、メモリ142、ホストインタフェース143、ディスクインタフェース144、キャッシュメモリ145、データコントローラ146を備えている。

【0028】

CPU141はディスクアレイ装置10の全体の制御を司る。CPU141はメモリ142に格納されているプログラムを実行することによりハードディスクドライブ120の管理やブロックアクセス要求の解釈など様々な機能を実現することができる。

【0029】

ホストインタフェース143は情報処理装置20との間で通信を行うインタフェースである。ホストインタフェース143はファイバチャネルプロトコルに従ってブロックアクセス要求を受け付ける機能を備える。

ディスクインタフェース144はハードディスクドライブ120との間でデータのやり取りを行うインタフェースである。ディスクインタフェース144はハードディスクドライブ120を制御するコマンドなどを規定するプロトコルに従ってハードディスクドライブ120に対するデータ入出力要求を送信する機能を備える。ディスクインタフェース144は、SCSIやファイバチャネルなどのプロトコルに従ってハードディスクドライブ120に対してデータの書き込みや読み出しのコマンドを送信することができる。

【0030】

キャッシュメモリ145は、ホストインタフェース143とディスクインタフェース144との間で授受されるデータが記憶されるメモリである。

データコントローラ146は、CPU141の制御によりホストインタフェース143とキャッシュメモリ145との間、あるいはキャッシュメモリ145とディスクインタフェース144との間でのデータ転送を行うものである。データコントローラ146は、例えば、特定用途向けICにロジックを形成する回路とすることができる。

【0031】

情報処理装置20がハードディスクドライブ120に対するデータの書き込み要求をディスクアレイ装置10に送信すると、ディスクアレイ装置10では、ホストインタフェース143が書き込み要求を受け付け、データコントローラ146が書き込み要求に付随する書き込みデータをキャッシュメモリ145に転送する。書き込みデータがキャッシュメモリ145に転送されると、データコントローラ146がキャッシュメモリ145からディスクインタフェース144に書き込みデータを読み出し、ディスクインタフェース144がハードディスクドライブ120に対して書き込みを指示するコマンドを送信する。

【0032】

なお、キャッシュメモリ145は不揮発とすることもできる。この場合、ホストインタフェース143が情報処理装置20から受信したデータがキャッシュメモリ145に書き込まれた段階で、CPU141が情報処理装置20に書込完了通知を送信することとしてもよい。

【0033】

以上、ディスクアレイ装置10の構成について説明した。なお、ディスクアレイ装置10は、以上に説明した構成のもの以外にも、例えば、NFS (Network File System) などのプロトコルにより情報処理装置20からファイル名指定によるデータ入出力要求を受け付けるように構成されたNAS (Network Attached Storage) として機能するものなどであってもよい。

【0034】

==データベースの運用形態==

次に、本実施の形態における、ディスクアレイ装置10に構築されたデータベースの運用形態について説明する。

図5は、クライアント端末30、データベースサーバ40、ディスクアレイ装置10で構成されるデータベースシステムを示す図である。データベースサーバ40は、図1および図4における情報処理装置20に相当する。

図6は、クライアント端末30の構成を示すブロック図である。クライアント端末30は、CPU601、メモリ602、記憶装置603、ポート604、記録媒体読取装置605、入力装置606、出力装置607を備える。

【0035】

CPU601はクライアント端末30の全体の制御を司るもので、メモリ602や記憶装置603に格納されたプログラムを実行することにより各種機能を実現する。記憶装置603は、例えばハードディスクドライブなどの記憶装置である。記録媒体読取装置60

5は、記録媒体608に記録されているプログラムやデータを読み取るための装置である。読み取られたプログラムやデータはメモリ602や記憶装置603に格納される。記録媒体608としてはフレキシブルディスクやCD-ROM、半導体メモリ等を用いることができる。記録媒体読取装置605はクライアント端末30に内蔵されている形態とすることもできるし、外付されている形態とすることもできる。入力装置606はオペレータ等によるクライアント装置30へのデータ入力等のために用いられる。入力装置606としては例えばキーボードやマウス等が用いられる。出力装置607は情報を外部に出力するための装置である。出力装置607としては例えばディスプレイやプリンタ等が用いられる。ポート604はデータベースサーバ40と通信を行うための装置である。

【0036】

図7は、データベースサーバ40の構成を示すブロック図である。データベースサーバ40は、CPU701、メモリ702、記憶装置703、LANインタフェース704、ストレージインタフェース705、記録媒体読取装置706を備える。

【0037】

CPU701はデータベースサーバ40の全体の制御を司るもので、メモリ702や記憶装置703に格納されたプログラムを実行することにより各種機能を実現する。記憶装置703は、例えばハードディスクドライブなどの記憶装置である。記録媒体読取装置706は、記録媒体707に記録されているプログラムやデータを読み取るための装置である。読み取られたプログラムやデータはメモリ702や記憶装置703に格納される。記録媒体707としてはフレキシブルディスクやCD-ROM、半導体メモリ等を用いることができる。記録媒体読取装置706はデータベースサーバ40に内蔵されている形態とすることもできるし、外付されている形態とすることもできる。

【0038】

LANインタフェース704はクライアント端末30とLAN等の通信手段により通信を行うためのインタフェースである。また、ストレージインタフェース705は、ストレージ装置10とSANやLAN等の通信手段により通信を行うためのインタフェースである。

【0039】

図5に示すように、ディスクアレイ装置10には、データ領域501とREDOログ領域502とが設けられている。データ領域501とREDOログ502領域は、一つ又は複数のハードディスクドライブ120により構成される記憶領域である。クライアント端末30は、データ領域501に格納されているテーブル503、504の更新要求をデータベースサーバ40に送信する。データベースサーバ40には、データベースバッファ505とREDOログバッファ506とが設けられている。データベースバッファ505とREDOログバッファ506は後述する記憶装置703に記憶されている。データベースバッファ505に格納されているデータは、データ領域501に格納されているデータのキャッシュデータとして使用される。これにより、データベースサーバ40はディスクアレイ装置10へのアクセスを行うことなく、クライアント端末30からのテーブル更新要求に応答することができ、クライアント端末30への応答時間を短縮することが可能となる。同様に、REDOログバッファ506に格納されているデータは、REDOログ領域502に格納されているデータのキャッシュデータとして使用される。

【0040】

データベースの更新処理においては、複数のテーブルの更新を保証することが要求される。例えば、銀行の預金口座間における振込の場合、振込元の口座を管理するテーブルに記憶されている預金残高を減額し、振込先の口座を管理するテーブルに記憶されている預金残高を増額する必要がある。つまり、振込元のテーブルの更新と振込先のテーブルの更新とが完了しないと、更新処理が正しく行われたことにならない。データベースサーバ40は、このように複数のテーブルの更新を保証する機能を備えている。クライアント端末30は、複数のテーブルの更新処理を実行した後に、これら複数のテーブルの更新処理の保証を要求するためのコミット要求をデータベースサーバ40に送信する。データベース

サーバ40は、これら複数のテーブルの更新処理の途中で何らかの障害が発生した場合は、これら複数のテーブルを更新前の状態に戻すことができる。

【0041】

図8は、データ領域501のテーブル503、504を更新する処理を示すフローチャートである。クライアント端末30は、データベースサーバ40にテーブル503の更新要求を送信する(S801)。更新要求にはテーブル503における更新位置と更新データが含まれている。更新要求を受信すると、データベースサーバ40は更新位置情報をもとに更新データをデータベースバッファ505に格納する(S802)。そして、データベースサーバ40は、更新位置情報と更新データとをREDOログバッファ506に格納し(S803)、クライアント端末30に更新完了通知を送信する(S804)。続いて、クライアント端末30はテーブル504の更新要求をデータベースサーバ40に送信する。データベースサーバ40は、同様にデータベースバッファ505及びREDOログバッファ506の更新を行い(S806、S807)、更新完了通知をクライアント端末30に送信する(S808)。テーブル504の更新完了通知を受信すると、クライアント端末30は、当該更新処理のコミット要求をデータベースサーバ40に送信する(S809)。

コミット要求を受信すると、データベースサーバ40は、REDOログバッファ506に格納されているテーブル503及びテーブル504の更新における更新位置情報と更新データとをREDOログ領域502に書き込む更新要求をディスクアレイ装置10に送信する(S810)。ディスクアレイ装置10はREDOログ領域502に更新位置情報と更新データとを格納し(S811)、更新完了通知をデータベースサーバ40に送信する(S812)。データベースサーバ40は、ディスクアレイ装置10からREDOログ領域502の更新完了通知を受信すると、コミット完了通知をクライアント端末30に送信する(S813)。クライアント端末30は、データベースサーバ40からコミット完了通知を受信し(S814)、テーブル503、504の更新が保証されたことを認識する。

【0042】

データベースバッファ505に格納された更新データは、クライアント端末30との間で行われる一連の処理(S801～S814)とは非同期に、ディスクアレイ装置10のデータ領域501に格納される。まず、データベースサーバ40は、データ領域501から更新前のデータの読み出す要求をディスクアレイ装置10に送信する(S851)。ディスクアレイ装置10は、更新前のデータをデータ領域501から読み出し、データベースサーバ40に送信する(S852)。データベースサーバ40は、更新前のデータと更新位置情報とをREDOログ領域502に書き込む要求をディスクアレイ装置10に送信する(S853)。ディスクアレイ装置10は、更新前のデータと更新位置情報とをREDOログ領域502に格納し(S854)、更新完了通知をデータベースサーバ40に送信する(S855)。データベースサーバ40は、REDOログ領域502の更新完了通知を受信すると、データベースバッファ505に格納されている更新データをデータ領域501に書き込む要求をディスクアレイ装置10に送信する(S856)。そして、ディスクアレイ装置10は更新データをデータ領域501に格納する(S857)。

このように、データベースサーバ40は、時間のかかるデータ領域501への更新データの格納を、データベースバッファ505への更新データの格納とは別のタイミングで行っている。これにより、データベースサーバ40は、クライアント端末30からの更新要求に対する応答時間を短縮している。

【0043】

また、データベースサーバ40に障害が発生した場合は、REDOログ領域502に格納されているデータ(以後、「REDOログ」と称する)を確認することにより、データ領域501のデータをクライアント端末30からのコミット要求を反映した内容に復旧することが可能である。つまり、REDOログに、S811の処理で格納されるコミット要求に伴う更新データが存在せず、かつ、S854の処理で格納される更新前データが存在

する場合、更新前データ領域501に書き込む。これを、ロールバック処理という。また、REDOログに、S811の処理で格納されるコミット要求に伴う更新データが存在し、かつ、S854の処理で格納される更新前データが存在しない場合、更新データをデータ領域501に書き込む。これを、ロールフォワード処理という。

このように、データ領域501のデータをクライアント端末30からのコミット要求を反映した内容に復旧する処理は、データベースのリカバリ処理と呼ばれている。リカバリ処理は、前述した手順に限られない。例えば、更新前データの全てをロールバックし、コミット要求に伴う更新データを全てロールフォワードすることにより実現されるものとしてもよい。

【0044】

次に、図9にて、複数台のハードディスクドライブ120に障害が発生し、RAIDの冗長性等を用いてデータ領域501のデータを復旧することができない場合（以後、「データ領域の障害」と称する）における、データ領域501の一般的な復旧手順について説明する。

【0045】

図9は、REDOログの記録を0時に開始し、12時にデータ領域501に障害が発生した場合を示している。また、データ領域501の6時の状態がデータ領域501を構成するハードディスクドライブ120とは別のハードディスクドライブ120又は磁気テープ等の記憶媒体にバックアップされている（S901）。なお、データ領域501のある時点のデータ（以後、「静止化データ」と称する）をバックアップする一般的な手順については、後述する。

【0046】

12時にデータ領域501の障害が発生した場合、障害が発生している一つ又は複数のハードディスクドライブ120を交換した後、バックアップされている6時のデータをデータ領域501に復元する（S902）。そして、前述したリカバリ処理の手順に従い、0時以降の更新前データの全てをロールバックし（S903）、0時から障害発生前までのコミット要求に伴う更新データをロールフォワードする（S904）。これにより、障害直前の状態までデータ領域501を復旧することができる。

【0047】

また、ユーザの操作ミスによる障害やソフトウェア障害によりデータ領域501が不正な状態となった場合も、静止化データとREDOログとを用いてデータ領域501を復旧することができる。ユーザの操作ミスによる障害とは、例えば、ファイルシステム上でデータ領域501をフォーマットしてしまうことによりデータ領域501を失ってしまう場合などである。また、ソフトウェア障害とは、アプリケーションプログラムの不具合等により、データ領域501に格納されているデータに不整合が生じてしまう場合などである。このような障害が発生した場合は、障害時刻をある程度特定し、REDOログを用いたロールフォワード処理を障害時刻の前までとする。これにより、障害前の状態のデータ領域501を復旧することができる。

【0048】

以上のように、データ領域501に障害が発生した場合は、静止化データとREDOログとを用いて、データ領域501を復旧することが可能である。また、静止化データをディスクアレイ装置10内に保持しておくことにより、磁気テープ等からのデータ復元を行わずに、迅速にデータ領域501の復旧を行うことも可能である。

【0049】

==データバックアップ方式==

次に、データのバックアップ方式について説明する。図10はレプリカ方式、図11はスナップショット方式を示す図である。

【0050】

まず、図10を用いてレプリカ方式について説明する。レプリカ方式では、データ領域501とは別の記憶領域としてレプリカボリューム1001が設けられている。レプリカ

ボリューム1001は一つ又は複数のハードディスクドライブ120により構成される。ディスク制御装置110は、データベースサーバ40からデータ領域501の更新要求を受信すると、データ領域501とレプリカボリューム1001の双方に更新データを書き込む。このように、双方に更新データを書き込んでいる状態を同期状態と称する。ディスク制御部110は、データベースサーバ40等からレプリカボリュームへ1001への更新データの書き込みを停止するスプリット指示命令を受信する。スプリット指示命令を受信すると、ディスク制御部110は、レプリカボリューム1001への更新データの書き込みを停止する。つまり、レプリカボリューム1001はスプリット指示命令を受信した時点における、データ領域501の静止化データである。

【0051】

ディスク制御装置110は、レプリカボリューム1001の更新を停止している間にデータ領域501に書き込まれた更新データの位置情報を、メモリ1003の差分セグメントビットマップ1004に記憶する。メモリ1003は、図1の共有メモリ133、キャッシュメモリ134、図4のキャッシュメモリ145などである。ディスク制御装置110は、データベースサーバ40等からレプリカボリューム1001を同期状態にするリシンク指示命令を受信すると、差分セグメントビットマップ1004に記憶されている位置情報で示されるデータをデータ領域501から読み出してレプリカボリューム1001に書き込む。このリシンク指示命令に伴う処理をリシンク処理と称する。

【0052】

更新を停止している間にデータ領域501に書き込まれたデータの量が多い場合、リシンク処理には相当の時間を要する。リシンク処理では、データ領域501におけるデータの更新順序とは関係なく、セグメント単位でレプリカボリューム1001のデータが更新される。つまり、リシンク処理の実行中は、レプリカボリューム1001はデータ領域501のある時点の状態とはなっていない。したがって、リシンク処理の実行中にデータ領域501に障害が発生した場合、レプリカボリューム1001をデータ領域501を復旧するために用いることができない。

【0053】

そこで、ディスクアレイ装置10内に、もう一つレプリカボリューム1002が設けられている場合もある。この場合、2つのレプリカボリューム1001、1002で交互に更新の停止とリシンク処理を行うことで、ディスクアレイ装置10内に必ずデータ領域501の静止化データを確保することができる。

【0054】

次に、図11を用いてスナップショット方式について説明する。スナップショット方式では、データ領域501とは別の記憶領域としてプール1101が設けられている。プール1101は一つ又は複数のハードディスクドライブ120により構成される。また、仮想的なレプリカボリュームである仮想レプリカボリューム1102が提供される。ディスク制御装置110は、データベースサーバ40等から仮想レプリカボリューム1102の作成指示を受信する。ディスク制御装置110は、データベースサーバ40からデータ領域501の更新要求を受信すると、データ領域501の更新予定位置に格納されている更新前データを読み出してプール1101に書き込む。そして、ディスク制御装置110は、データ領域501に更新データを書き込み、メモリ1003の差分セグメントビットマップ1004に更新データの位置情報を記憶する。ディスク制御装置110は、仮想レプリカボリューム1102に対する読み出し要求を受信すると、差分セグメントビットマップ1004を参照し、読み出しデータの位置情報が記憶されている場合はプール1101からデータを読み出し、そうでない場合はデータ領域501からデータを読み出す。つまり、仮想レプリカボリューム1102は、仮想レプリカボリューム1102の作成指示を受信した時点における、データ領域501の静止化データである。

【0055】

以上に説明したレプリカ方式とスナップショット方式には、次のような欠点がある。レプリカ方式では、リシンク処理の際にデータ領域501に格納されているデータを読み出

す必要がある。そのため、データベースサーバ40から受信するデータ領域501の更新要求に伴うデータ領域501の更新処理と競合し、データベースサーバ40からデータ領域501へのアクセス性能が低下してしまう。また、2つのレプリカボリューム1001、1002を設ける場合、データ領域501の3倍の記憶容量が必要となり、ディスクアレイ装置10の導入コストが高くなってしまう。

【0056】

スナップショット方式では、静止化データである仮想レプリカボリューム1102を読み出す際に、データ領域501に格納されているデータを読み出す必要がある。そのため、レプリカ方式の場合と同様、データベースサーバ40から受信するデータ領域501の更新要求に伴うデータ領域501の更新処理と競合し、データベースサーバ40からデータ領域501へのアクセス性能が低下してしまう。また、データ領域501において障害が発生した場合、ディスクアレイ装置10内にデータ領域501の静止化データが存在しないこととなる。そのため、データ領域501を迅速に復旧することができない。

【0057】

次に、前述したレプリカ方式とスナップショット方式を改良した、本実施の形態におけるデータバックアップ方式を、図12を用いて説明する。

【0058】

ディスクアレイ装置10には、データ領域501が格納される第一の記憶デバイス1201、第二の記憶デバイス1202、及び第三の記憶デバイス1203が設けられている。第一乃至第三の記憶デバイス1201～1203は、それぞれ一つ又は複数のハードディスクドライブ120により構成されている。また、第一の記憶デバイス1201及び第二の記憶デバイス1202には、一つ又は複数の論理ボリューム121が形成されている。また、第一の記憶デバイス1201を構成するハードディスクドライブ120と第二の記憶デバイス1202及び第三の記憶デバイス1203を構成するハードディスクドライブ120は物理的に別のものである。なお、REDOログは、ディスクアレイ装置10内の第一の記憶デバイス1201を構成するハードディスクドライブ120とは別のハードディスクドライブ120に格納されている。また、REDOログは、ディスクアレイ装置10の外部にある記憶装置に格納されているものとしてもよい。ディスク制御装置110には、計時機構1204が備えられている。計時機構1204は、外部からの要求に応じて日付及び時刻を提供する。計時機構1204は、クロック信号により時刻をカウントする計時回路などである。メモリ1003には、所定の時間である検出時間1205が記憶されている。検出時間1205は、管理端末136や情報処理装置20などから登録される。

【0059】

本方式における、データの更新処理の概略を説明する。ディスク制御装置110は、データベースサーバ40から第一の記憶デバイス1201の論理ボリューム121の更新要求を受信する。ディスク制御装置110は、当該更新要求を受信すると、第一の記憶デバイス1201の論理ボリュームに更新データを書き込む。そして、ディスク制御部110は第一の記憶デバイスに書き込んだ更新データと更新データが書き込まれた位置情報、更新時刻等で構成されるジャーナルデータを第三の記憶デバイスに書き込む。ディスク制御部110は、第三の記憶デバイスに書き込まれているジャーナルデータを所定の間隔で参照し、更新時刻と現在時刻との差が検出時間1205を超えているジャーナルデータについて、更新時刻順に更新データを第二の記憶デバイス1202の論理ボリューム121に書き込む。第二の記憶デバイス1202の論理ボリューム1202は、第一の記憶デバイス1201の論理ボリューム121の検出時間1205だけ前の状態となっている。つまり、第一の記憶デバイス1201に障害が発生した場合、第二の記憶デバイス1202と第三の記憶デバイス1203又はREDOログとを用いて、第一の記憶デバイス1201を復旧することが可能である。

【0060】

図13は、本実施の形態に係るディスク制御装置110における前述したデータの更新

処理を実現する機能を示すブロック図である。ディスク制御装置110は、書き込み要求受信部1301、第一の書き込み部1302、ジャーナル書き込み部1303、第二の書き込み部1304、スプリット指示命令受信部1305、スプリット解除命令受信部1306、スプリット指示記憶部1307、スプリット解除記憶部1308、未反映情報記憶部1309、読み出し要求受信部1310、識別子取得部1311、重複範囲取得部1312、仮想論理ボリューム読み出し部1313、読み出しデータ送信部1314を備えている。各部1301~1314は、図1~図3に示すディスクアレイ装置10におけるCPU203及びCPU303、図4に示すディスクアレイ装置10におけるCPU141がメモリ202、302、142又はNVRAM204、304に格納されているプログラムを実行することにより実現される。

【0061】

==ペア及びグループ==

図14は、第一の記憶デバイス1201の論理ボリューム121と第二の記憶デバイス1202の論理ボリューム121との関係を示す図である。第一の記憶デバイス1201の論理ボリューム121は、第二の記憶デバイス1202の論理ボリューム121と対応付けられている。この対応付けにおける第一の記憶デバイス1201の論理ボリューム121を主論理ボリューム、第二の記憶デバイス1202の論理ボリューム121を副論理ボリュームと称する。この対応付けはペアと呼ばれ、図15に示すペア管理テーブル1501に主論理ボリュームの識別子と副論理ボリュームの識別子とが対応付けられて記憶されている。また、第一の記憶デバイス1201の論理ボリューム121は一つ又は複数の論理ボリューム121をまとまりとするグループを形成している。グループは、例えば、データベースサーバ40ごとや、データベースサーバ40が提供するデータベース・インスタンスごとに形成される。つまり、データベースサーバ40は第一の記憶デバイス1201の論理ボリューム121をグループ単位で使用していると言うことができる。従って、第一の記憶デバイス1201の論理ボリューム121をグループ単位でバックアップすることが必要となる。グループにはグループを示すグループIDが付与され、図16に示すグループ管理テーブル1601に論理ボリューム121の識別子とグループIDとの対応付けが記憶されている。ペア管理テーブル1501及びグループ管理テーブル1601はメモリ1003に記憶されており、管理端末136や情報処理装置20などから登録される。

【0062】

==ジャーナルデータ==

図17は、第三の記憶デバイス1203に書き込まれるジャーナルデータを示す図である。ジャーナルデータ1701は、ジャーナル部1702及び更新ストリーム部1703からなる。

【0063】

ジャーナル部1702は、グループID、シーケンスID、更新時刻、更新位置情報、及び更新ストリームオフセットで構成されている。データベースサーバ40から第一の記憶デバイス1201の論理ボリューム121に更新データを書き込む要求を受信すると、ディスク制御装置110は第三の記憶デバイス1203にジャーナルデータ1701を記憶する。グループIDは、当該論理ボリューム121が属するグループのグループIDである。シーケンスIDは、第一の記憶デバイス1201における更新データの更新順序を示すものであり、グループごとに管理されている連続番号である。ジャーナルデータ1701に格納されている更新データは、グループごとにシーケンスID順に第二の記憶デバイス1202の論理ボリューム121に書き込まれる。更新時刻は、ディスク制御装置110がジャーナルデータ1701を記憶する際に計時機構1201から取得した日付と時刻である。したがって、シーケンスIDの順序と更新時刻の順序は同じである。更新位置情報には、当該論理ボリューム121の識別子と当該論理ボリューム121において更新データが格納された位置情報とが記憶されている。なお、データが格納された位置情報とは、例えば、開始アドレス及び開始アドレスからのデータ長などで示される情報である。

更新ストリームオフセットには、更新データの更新ストリーム部1703における位置情報が記憶されている。更新ストリーム部1703には、更新データが更新ストリームオフセットで示される位置に記憶されている。

【0064】

また、ディスク制御部110が後述するスプリット指示命令を受信した場合は、更新位置情報にスプリット指示命令を受信したことを示すデータが記憶される。グループIDには、スプリット指示命令で指定されたグループIDが記憶される。シーケンスIDには、スプリット指示命令で指定されたグループIDのシーケンスIDが記憶される。更新時刻には、計時機構1201から取得した日付と時刻が記憶される。このように、スプリット指示命令に伴いジャーナル部1701に書き込まれたデータをスプリット指示データと称する。そして、ディスク制御部110が後述するスプリット解除命令を受信した場合は、更新位置情報にスプリット解除命令を受信したことを示すデータが記憶される。このように、スプリット解除命令に伴いジャーナル部1701に書き込まれたデータをスプリット解除データと称する。

【0065】

なお、第二の記憶デバイス1202の論理ボリューム121への反映が完了した更新データを示すジャーナルデータ1701の記憶領域は空き領域として管理され、別の更新データを示すジャーナルデータ1701の記憶領域として再利用される。

【0066】

==スプリット運用==

次に、本実施の形態において静止化データを作成してデータのバックアップを行う概要について説明する。図18は、静止化データを作成するためのスプリット指示命令を受信した場合の、ディスクアレイ装置10の状態遷移を示す図である。

【0067】

まず、通常運用の状態(S1801)においては、第三の記憶デバイス1203のジャーナルデータ1701から、更新時刻と現在時刻との差が検出時間1205を超えている更新データが抽出され、第二の記憶デバイス1202に書き込まれている。

【0068】

ディスクアレイ装置10は、データベースサーバ40またはバックアップサーバ1803等からスプリット指示命令を受信すると、仮想デバイス1801を作成する状態(S1802)へと遷移する。バックアップサーバ1803とは、図1及び図4における情報処理装置10に相当する装置である。バックアップサーバ1803は、ディスクアレイ装置10に格納されているデータを読み出し、外部の記憶媒体にバックアップする機能を備えている。

【0069】

図18のS1802は、ディスクアレイ装置10がスプリット指示命令を18時(スプリット時刻)に受信してから5分経過した状態を示している。なお、図18の例では検出時間は2時間である。S1802において、第二の記憶デバイス1202には、現在時刻の2時間前である16:05までの更新データが反映されている。つまり、16:05から18:00までの更新データは、まだ第二の記憶デバイス1202に反映されず、第三の記憶デバイス1203に格納されている。

【0070】

そこで、18:00における第一の記憶デバイスの静止化データとして仮想デバイス1801が提供される。仮想デバイス1801には、仮想的な論理ボリュームである仮想論理ボリュームが設けられている。仮想論理ボリュームの識別子と第二の記憶デバイス1202の論理ボリューム121の識別子との対応は、図19に示す仮想論理ボリューム管理テーブル1901に記憶されている。仮想論理ボリューム管理テーブル1901はメモリ1003に記憶され、管理端末136や情報処理装置20などから登録される。

【0071】

ディスク制御装置110は、更新時刻がスプリット時刻より早いジャーナルデータ17

01の更新位置情報を抽出し、未反映情報としてメモリ1003に記憶している。未反映情報の記憶方式は、例えば、第二の記憶デバイス1202の記憶領域を複数のセグメントに分割し、このセグメント単位に未反映の更新データがあるかどうかを示すビットマップ方式などである。ディスク制御装置110は、バックアップサーバ1802から仮想論理ボリュームに対するデータの読み出し要求を受信すると、未反映情報で示される位置に格納されているデータについては、第三の記憶デバイス1203からデータを読み出す。また、ディスク制御装置110は、未反映情報で示される位置以外に格納されているデータについては、第二の記憶デバイス1202から読み出す。これにより、バックアップサーバ1802に対して第一の記憶デバイス1201の18時の静止化データを提供することができる。

【0072】

仮想デバイス1801が提供されている間も、第三の記憶デバイス1203から第二の記憶デバイス1202への更新データの反映は行われている。そのため、ディスク制御装置110は、仮想デバイス1801を静止化データとするために、更新時刻がスプリット時刻以降の更新データを第二の記憶デバイス1202に反映しない。この状態を示しているのが、S1803である。S1803においては、第二の記憶デバイス1202にはスプリット時刻である18時までの更新データが反映されている。したがって、第二の記憶デバイス1202と仮想デバイス1801とは一致している。

【0073】

ディスクアレイ装置10は、データベースサーバ40またはバックアップサーバ1801等からスプリット解除命令を受信すると、第三の記憶デバイス1203に格納されている、更新時刻と現在時刻との差が検出時間1205を超えている更新データの第二の記憶デバイス1202への反映を再開する。S1804は、この状態を示している。そして、この反映が完了すると、通常運用の状態(S1801)に戻る。

【0074】

なお、仮想デバイス1801を用いずに、スプリット指示命令を受信するとディスク制御装置110は更新時刻がスプリット時刻より早い更新データを、更新時刻と現在時刻との差が検出時間1205を超えているかどうかにかかわらず、第二の記憶デバイス1202に反映するものとしてもよい。この場合、バックアップサーバ1802は、仮想デバイス1801ではなく、第二の記憶デバイス1202からデータを読み出してバックアップを行う。

【0075】

==動作の説明==

以上に説明した、本実施の形態におけるバックアップ方式を実現するための、ディスク制御装置110の各部1301~1314の動作を説明する。

【0076】

==第三の記憶デバイスへの書き込み==

図20は、ディスク制御装置110がデータベースサーバ40から第一の記憶デバイス1201の論理ボリューム121に対するデータの書き込み要求を受信した際の処理を示すフローチャートである。

【0077】

書き込み要求受信部1301は、データベースサーバ40から第一の記憶デバイス1201の論理ボリューム121に対するデータの書き込み要求と書き込み用のデータとを受信する(S2001)。第一の書き込み部1302は、当該データを第一の記憶デバイス1201の当該書き込み要求で指定される論理ボリューム121に書き込む(S2002)。ジャーナル書き込み部1303は、当該データに対するジャーナルデータ1701を第三の記憶デバイス1203に書き込む(S2003)。そして、ジャーナル書き込み部1303は、データベースサーバ40に対して当該データの書き込み完了通知を送信する(S2004)。

【0078】

なお、当該ジャーナルデータ1701の生成は次のように行われる。ジャーナル書き込み部1303は、当該論理ボリュームの識別子に対応するグループIDをグループ管理テーブル1601から取得し、当該ジャーナルデータ1701のグループIDに設定する。ジャーナル書き込み部1303は、取得したグループIDにおいて前回付与されたシーケンスIDに1加算し、当該ジャーナルデータ1701のシーケンスIDに設定する。ジャーナル書き込み部1303は、計時機構1204から現在の日時を取得し、当該ジャーナルデータ1701の更新時刻に設定する。ジャーナル書き込み部1303は、当該論理ボリュームの識別子と当該データが書き込まれた位置情報とを当該ジャーナルデータ1701の更新位置情報に設定する。ジャーナル書き込み部1303は、更新ストリーム部1703において次にデータが格納される位置を当該ジャーナルデータ1701の更新ストリームオフセットに設定する。そして、ジャーナル書き込み部1303は、当該データを更新ストリーム部1703に格納する。

【0079】

また、ジャーナルデータ1701は、第三の記憶デバイス1203の論理ボリューム121に書き込まれる前に、メモリ1003に書き込まれる。メモリ1003が冗長構成で、かつ不揮発性である場合、メモリ1003にジャーナルデータ1701が書き込まれた時点でジャーナル書き込み部1303がデータベースサーバ40に対して当該データの書き込み完了通知を送信することとしてもよい。

【0080】

==スプリット指示==

図21は、ディスク制御装置110がスプリット指示命令を受信した際の処理を示すフローチャートである。

【0081】

スプリット指示命令受信部1305は、データベースサーバ40やバックアップサーバ1801等からスプリット指示命令を受信する(S2101)。スプリット指示記憶部1307は、スプリット指示データを第三の記憶デバイス1203に書き込む(S2102)。なお、スプリット指示データの生成は次のように行われる。スプリット指示記憶部1307は、スプリット指示命令で指定されるグループIDを当該スプリット指示データのグループIDに設定する。スプリット指示記憶部1307は、当該グループIDにおいて前回付与されたシーケンスIDに1加算し、当該スプリット指示データのシーケンスIDに設定する。スプリット指示記憶部1307は、計時機構1204から現在の日時を取得し、当該スプリット指示データの更新時刻に設定する。スプリット指示記憶部1307は、スプリット指示命令を受信したことを示すデータを当該スプリット指示データの更新位置情報に設定する。

そして、未反映情報記憶部1309は、更新時刻がスプリット時刻より早いジャーナルデータ1701の更新位置情報である未反映情報をメモリ1003に記憶する。なお、更新時刻がスプリット時刻より早いかどうかの判断は、更新時刻を直接比較することにより行われてもよいし、シーケンスIDを比較することにより行われてもよい。

【0082】

==スプリット解除==

図22は、ディスク制御装置110がスプリット解除命令を受信した際の処理を示すフローチャートである。

スプリット解除命令受信部1306は、データベースサーバ40やバックアップサーバ1801等からスプリット解除命令を受信する(S2201)。スプリット解除記憶部1308は、スプリット解除データを第三の記憶デバイス1203に書き込む(S2202)。つまり、スプリット解除記憶部1308は、スプリット指示命令で指定されるグループIDが設定されているスプリット指示データの更新位置情報にスプリット解除命令を受信したことを示すデータを設定する。

【0083】

==第二の記憶デバイスへの更新データの反映==

図23は、ディスク制御装置110が第三の記憶デバイス1203に格納されている更新データを第二の記憶デバイス1202に反映する処理を示すフローチャートである。

第二の書き込み部1304は、メモリ1003に格納されている検出時間1205を取得する(S2301)。第二の書き込み部1304は、計時機構1204から日付と時刻(現在時刻)を取得する(S2302)。第二の書き込み部1304は、第三の記憶デバイス1203に格納されているジャーナルデータ1701を参照し(S2303)、スプリット指示データがあり、かつ、スプリット解除データがないかどうかを確認する(S2304)。

【0084】

スプリット指示データがあり、かつ、スプリット解除データがない場合、第二の書き込み部1304は、当該スプリット指示データとグループIDが同一で、更新時刻と現在時刻との差が検出時間1205を超えており、更新時刻がスプリット時刻より早いジャーナルデータを選択する(S2305)。条件に該当するジャーナルデータ1701がない場合は、再びS2302からの処理を実行する。なお、更新時刻がスプリット時刻より早いジャーナルの選択は、更新時刻を比較することにより行われてもよいし、シーケンスIDを比較することにより行われてもよい。

スプリット指示データがないか、またはスプリット解除データがある場合、第二の書き込み部1304は、更新時刻と現在時刻との差が検出時間1205を超えているジャーナルデータ1701を選択する(S2306)。条件に該当するジャーナルデータ1701がない場合は、再びS2302からの処理を実行する。

【0085】

第二の書き込み部1304は、選択したジャーナルデータ1701の更新位置情報に設定されている論理ボリューム121の識別子に対応する副論理ボリュームの識別子をペア管理テーブル1501から取得する(S2307)。第二の書き込み部1304は、選択したジャーナルデータ1701の更新データを、取得した副論理ボリュームの識別子で示される第二の記憶デバイス1202の論理ボリューム121のジャーナルデータ1701の位置情報で示される場所書き込む(S2308)。なお、ジャーナルデータ1701が複数ある場合、第二の書き込み部1304は、シーケンスID順、つまり更新時刻の早い順に更新データを第二の記憶デバイス1202の論理ボリューム121に書き込む。第二の記憶デバイス1202の論理ボリューム121への書き込みが完了すると、第二の書き込み部1304は、メモリ1003に記憶されている当該更新データの未反映情報を削除する(S2309)。

【0086】

なお、前述のS2305のステップを、第二の書き込み部1304が更新時刻と現在時刻との比較を行わず、更新時刻がスプリット時刻より早いジャーナルデータを選択すると変更することにより、仮想デバイス1801を設けずに第二の記憶デバイス1202を第一の記憶デバイス1201の静止化データとすることができる。

【0087】

==仮想論理ボリュームの読み出し==

図24は、ディスク制御装置110が仮想デバイス1801の仮想論理ボリュームに対する読み出し要求を受信した際の処理を示すフローチャートである。

【0088】

読み出し要求受信部1310は、バックアップサーバ1802から仮想デバイス1801の仮想論理ボリュームの識別子が指定された読み出し要求を受信する(S2401)。識別子取得部1311は、読み出し要求で指定された仮想論理ボリュームの識別子に対応する論理ボリューム121の識別子を仮想論理ボリューム管理テーブル1901から取得する(S2402)。重複範囲取得部1312は、読み出し要求で指定される位置情報とメモリ1003に記憶されている未反映情報の位置情報とで重複している範囲を取得する(S2403)。未反映情報が開始アドレスとブロック長とで示されている場合、重複している範囲とは、読み出し要求で指定される位置情報で示される範囲と未反映情報の位置

情報で示される範囲とで重なりあう部分である。また、未反映情報がセグメント単位でのビットマップで表現されている場合、重複している範囲とは、読み出し要求で指定される位置情報で示される範囲のうちで未反映情報のビットマップで未反映のビットが設定されているセグメントである。

【0089】

仮想論理ボリューム読み出し部1313は、重複している範囲については第三の記憶デバイス1203に格納されているジャーナルデータ1701から読み出し（S2404）、重複していない範囲については第二の記憶デバイス1202の論理ボリューム121から読み出す（S2405）。そして、読み出しデータ送信部1314は、読み出されたデータをバックアップサーバ1802に送信する（S2406）。

【0090】

以上、本実施の形態におけるバックアップ方式を実現するための、ディスク制御装置110の各部1301～1314の動作を説明した。

【0091】

これにより、第二の記憶デバイス1202の論理ボリューム121を第一の記憶デバイス1201の論理ボリューム121の検出時間1205前の状態とすることができる。ソフトウェアの障害等により不正なデータが第一の記憶デバイス1201の論理ボリューム121に書き込まれた場合、検出時間1205の間は、第二の記憶デバイス1202の論理ボリューム121には不正なデータが書き込まれない。

【0092】

例えば、情報処理装置20がデータベースサーバ40であり、第一の記憶デバイス1201の論理ボリューム121がデータベースのデータ格納領域であるとする。このような場合、データベースサーバ40は第一の記憶デバイス1201の論理ボリューム121に書き込んだデータに対するある時点からのREDOログを別の記憶領域に記憶している。したがって、第一の記憶デバイス1201の論理ボリューム121に不正なデータが書き込まれた場合、第二の記憶デバイス1202の論理ボリューム121に記憶されている検出時間1205前のデータとREDOログとを用いて、第一の記憶デバイス1201の論理ボリューム121を不正なデータが書き込まれる直前の状態にまで復元することができる。つまり、磁気テープ等の媒体に保存されているデータを用いることなく、ディスクアレイ装置10内のデータを用いて第一の記憶デバイス1201の論理ボリューム121のデータを復元することができるため、復元作業に要する時間を短縮することが可能である。

【0093】

また、第一の記憶デバイス1201の論理ボリューム121を主論理ボリュームとし、主論理ボリュームに対する副論理ボリューム（レプリカボリューム）を2つ設ける従来のデータバックアップ方式と比較して、記憶容量を削減することが可能である。副論理ボリュームを2つ設ける従来のデータバックアップ方式では、主論理ボリュームと副論理ボリュームとで、第一の記憶デバイス1201の3倍の記憶容量が必要であった。本発明の方式で必要となる記憶容量は、第三の記憶デバイス1203に格納するジャーナルデータの量に依存している。通常、1日のデータベース更新量は主論理ボリュームの20%までであることが多い。つまり、検出時間1205を1日とする場合、第三の記憶デバイス1203の記憶容量は第一の記憶デバイス1201の25%を確保すれば十分である。したがって、第一の記憶デバイス1201と第二の記憶デバイス1202と第三の記憶デバイス1203とで、第一の記憶デバイス1201の2.25倍の記憶容量でよいこととなる。検出時間1205を数時間とすることができる場合は、必要な記憶容量を更に削減することが可能である。つまり、障害が発生した場合にディスクアレイ装置10内に存在するデータで第一の記憶デバイス1201の論理ボリューム121を復旧するバックアップ方式において、必要な記憶容量を削減でき、ディスクアレイ装置10の導入コストを削減することができる。

【0094】

また、検出時間 1205 がメモリ 1003 に記憶されていることにより、検出時間 1205 を管理端末 136 や情報処理装置 20 等から登録することが可能となる。つまり、ソフトウェアの不具合や人為的な操作ミス等により発生する第一の記憶デバイス 1201 の論理ボリューム 121 に不正なデータが書き込まれていることを検出可能な時間を、業務やソフトウェアの特性に応じて変更することが可能である。したがって、第三の記憶デバイス 1203 にジャーナルデータ 1701 を格納するために必要な記憶容量を、業務やソフトウェアの特性に応じて変更することができる。

【0095】

また、第一の記憶デバイス 1201 を構成するハードディスクドライブ 120 と第二の記憶デバイス 1202 を構成するハードディスクドライブ 120 とを物理的に別のものとすることができる。第一の記憶デバイス 1201 を構成するハードディスクドライブ 120 にハードウェア障害が発生した場合でも、第二の記憶デバイス 1202 には、検出時間 1205 前に第一の記憶デバイス 1201 の論理ボリューム 121 に記憶されていたデータが記憶されている。情報処理装置 20 がデータベースサーバ 40 で、ディスクアレイ装置 10 内に第一の記憶デバイス 1201 の論理ボリューム 121 に書き込まれた REDO ログが存在する場合、第二の記憶デバイス 1202 の論理ボリューム 121 と REDO ログとで、第一の記憶デバイス 1201 の論理ボリューム 121 のデータを復元することができる。データの復元には、磁気テープ等の媒体に保存されているデータを用いないため、復元作業に要する時間を短縮することが可能である。

【0096】

つまり、第一の記憶デバイス 1201 の論理ボリューム 121 のある時点の静止化データを用いることなく、継続的に更新が行われている第二の記憶デバイス 1201 を用いて第一の記憶デバイス 1201 の論理ボリューム 121 のデータを迅速に復旧することができる。

【0097】

また、第一の記憶デバイス 1201 を構成するハードディスクドライブ 120 と第二の記憶デバイス 1202 及び第三の記憶デバイス 1203 を構成するハードディスクドライブ 120 とを物理的に別のものとすることができる。第三の記憶デバイス 1203 に格納されているジャーナルデータ 1701 を用いて第二の記憶デバイス 1202 の論理ボリューム 121 を更新する際に、第一の記憶デバイス 1201 からデータを読み出す必要がない。つまり、情報処理装置 20 から第一の記憶デバイス 1201 に対するデータ入出力処理の性能に与える影響が少ない。

【0098】

また、第一の記憶デバイス 1201 を構成するハードディスクドライブ 120 にハードウェア障害が発生した場合、第二の記憶デバイス 1202 の論理ボリューム 121 と第三の記憶デバイス 1203 に格納されているジャーナルデータ 1701 とを用いて第一の記憶デバイス 1201 の論理ボリューム 121 のデータを復元することができる。この場合においても、データの復元には磁気テープ等の媒体に保存されているデータを用いないため、復元作業に要する時間を短縮することが可能である。

【0099】

また、スプリット時刻から検出時間を経過後、第一の記憶デバイス 1201 のスプリット時刻の状態を第二の記憶デバイス 1202 上に作成することが可能となる。この状態において、第二の記憶デバイス 1202 の論理ボリューム 121 を磁気テープ等の記憶媒体にバックアップすることが可能となる。したがって、ディスクアレイ装置 10 全体の障害や広域災害等に備えたデータのバックアップを行うことが可能となる。

【0100】

また、スプリット指示後直ちに、第二の記憶デバイス 1202 の論理ボリューム 121 の仮想論理ボリュームを用いて、第一の記憶デバイス 1201 の論理ボリューム 121 のスプリット時刻におけるデータを磁気テープ等にバックアップすることが可能である。つまり、第三の記憶デバイス 1203 に格納されている更新時刻がスプリット時刻より早い

ジャーナルデータが第二の記憶デバイス1202の論理ボリューム121に反映されるのを待つ必要がない。そのため、データのバックアップを容易に行うことが可能となる。

【0101】

また、スプリット指示命令を受信すると、検出時間にかかわらず、更新時刻がスプリット時刻より早いジャーナルデータの更新データを第二の記憶デバイス1202の論理ボリューム121に反映することもできる。これにより、第二の記憶デバイス1202の論理ボリューム121を用いて、第一の記憶デバイス1201の論理ボリューム121のスプリット時刻におけるデータを磁気テープ等にバックアップすることが可能である。つまり、バックアップ処理中に仮想論理ボリュームを形成する必要がないため、ディスクアレイ装置10の負荷が軽減され、バックアップ処理時間が短縮される。

【0102】

また、スプリット指示命令を第一の記憶デバイス1201の論理ボリューム121のグループIDを指定して行うことができる。第一の記憶デバイス1201を、複数の情報処理装置20が使用している場合や、情報処理装置10上で複数の業務アプリケーションが稼働している場合等に、情報処理装置20や業務アプリケーションごとに論理ボリューム121にグループIDを付与することで、グループID単位でデータのバックアップを行うことが可能となる。

【0103】

例えば、情報処理装置20がデータベースサーバであり、データベースサーバが使用しているデータベース・インスタンスが2つあるとする。各データベース・インスタンスがデータ格納領域として用いている第一の記憶デバイス1201の論理ボリューム121に、異なるグループIDを付与する。スプリット指示命令及びスプリット解除命令は、このグループID単位で行われる。つまり、データベース・インスタンス単位でデータのバックアップを行うことが可能となる。また、例えば、業務アプリケーションごとに使用する論理ボリューム121が分かれている場合、業務アプリケーション単位でのデータのバックアップを行うことが可能となる。

【0104】

また、グループIDを考慮する場合においても、当該グループIDに対する仮想論理ボリュームを用いて、第一の記憶デバイス1201の論理ボリューム121のスプリット時刻におけるデータを磁気テープ等にバックアップすることが可能である。つまり、バックアップを開始する前に、第三の記憶デバイスに格納されている更新時刻がスプリット時刻より早いジャーナルデータが第二の記憶デバイスの論理ボリュームに反映されるのを待つ必要がない。そのため、データのバックアップを容易に行うことが可能となる。

【0105】

また、グループIDを考慮する場合においても、スプリット指示命令を受信すると、当該グループIDについては検出時間にかかわらず更新時刻がスプリット時刻より早いジャーナルデータの更新データを第二の記憶デバイス1202の論理ボリューム121に反映することもできる。これにより、第二の記憶デバイス1202の論理ボリューム121を用いて、第一の記憶デバイス1201の論理ボリューム121のスプリット時刻におけるデータを磁気テープ等にバックアップすることが可能である。つまり、バックアップ処理中に仮想論理ボリュームを形成する必要がないため、ディスクアレイ装置10の負荷が軽減され、バックアップ処理時間が短縮される。

【0106】

以上、本実施の形態について説明したが、上記実施例は本発明の理解を容易にするためのものであり、本発明を限定して解釈するためのものではない。本発明は、その趣旨を逸脱することなく、変更、改良され得ると共に、本発明にはその等価物も含まれる。

【図面の簡単な説明】

【0107】

【図1】本実施の形態に係るディスクアレイ装置の第一の形態の構成を示すブロック図である。

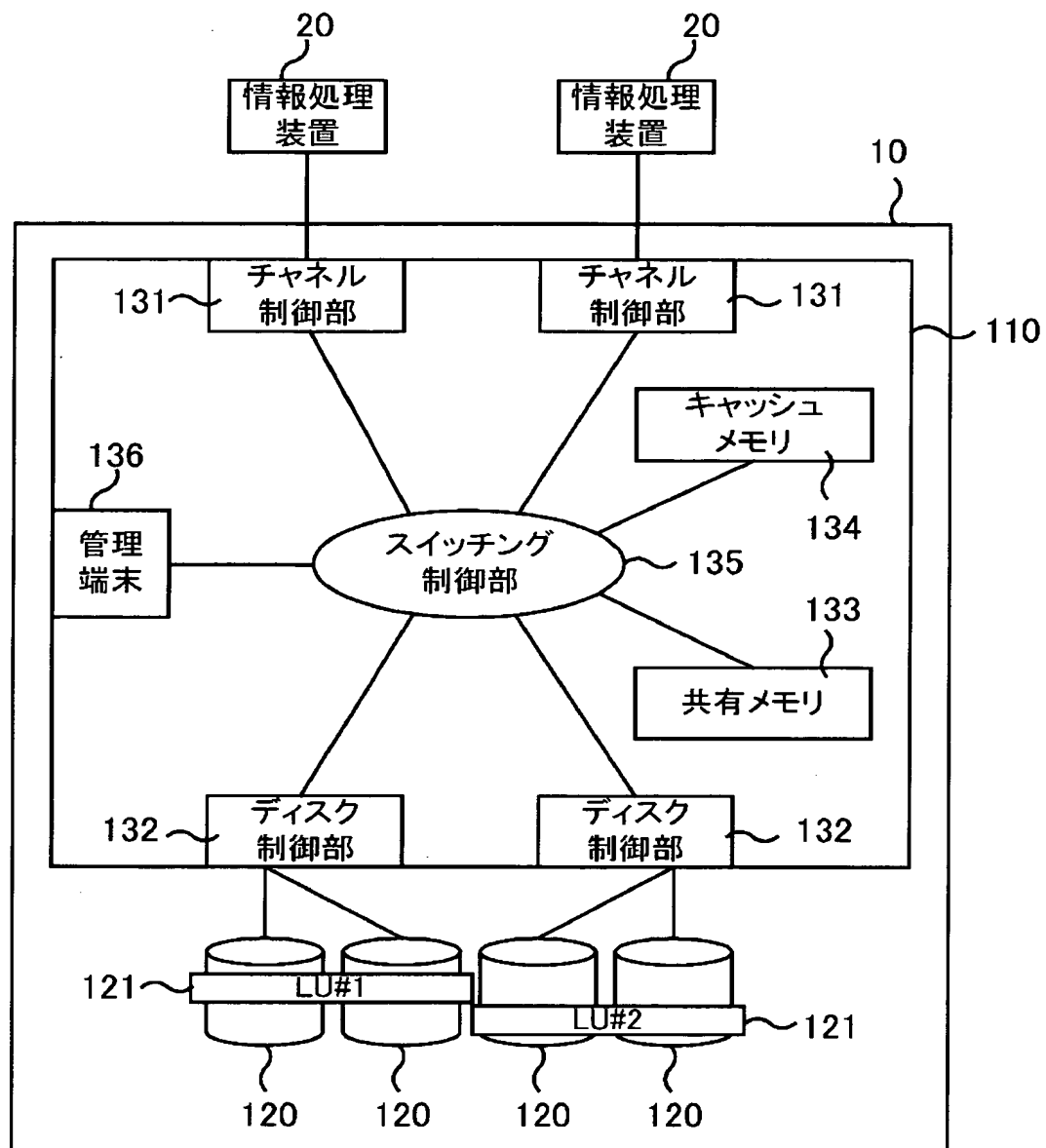
- 【図 2】本実施の形態に係るチャネル制御部の構成を示すブロック図である。
- 【図 3】本実施の形態に係るディスク制御部の構成を示すブロック図である。
- 【図 4】本実施の形態に係るディスクアレイ装置の第二の形態の構成を示すブロック図である。
- 【図 5】本実施の形態に係るクライアント端末、データベースサーバ、ディスクアレイ装置で構成されるデータベースシステムを示す図である。
- 【図 6】本実施の形態に係るクライアント端末の構成を示すブロック図である。
- 【図 7】本実施の形態に係るデータベースサーバの構成を示すブロック図である。
- 【図 8】本実施の形態に係るデータベースシステムにおけるテーブル更新処理を示すフローチャートである。
- 【図 9】REDO ログを用いてデータを復旧する一般的な手順を示す図である。
- 【図 10】レプリカボリュームを用いてデータのバックアップを行う方式を示す図である。
- 【図 11】スナップショットを用いてデータのバックアップを行う方式を示す図である。
- 【図 12】本実施の形態に係るバックアップ方式を示す図である。
- 【図 13】本実施の形態に係るバックアップ方式を実現する機能を示すブロック図である。
- 【図 14】本実施の形態に係る第一の記憶デバイスの論理ボリュームと第二の記憶デバイスの論理ボリュームとの関係を示す図である。
- 【図 15】本実施の形態に係るペア管理テーブルを示す図である。
- 【図 16】本実施の形態に係るグループ管理テーブルを示す図である。
- 【図 17】本実施の形態に係るジャーナルデータを示す図である。
- 【図 18】本実施の形態に係るスプリット指示命令を受信した場合のディスクアレイ装置の状態遷移を示す図である。
- 【図 19】本実施の形態に係る仮想論理ボリューム管理テーブルを示す図である。
- 【図 20】本実施の形態に係るディスク制御装置が第一の記憶デバイスの論理ボリュームに対する書き込み要求を受信した際の処理を示すフローチャートである。
- 【図 21】本実施の形態に係るディスク制御装置がスプリット指示命令を受信した際の処理を示すフローチャートである。
- 【図 22】本実施の形態に係るディスク制御装置がスプリット解除命令を受信した際の処理を示すフローチャートである。
- 【図 23】本実施の形態に係るディスク制御装置がジャーナルデータを第二の記憶デバイスに反映する処理を示すフローチャートである。
- 【図 24】本実施の形態に係るディスク制御装置が仮想論理ボリュームに対する読み出し要求を受信した際の処理を示すフローチャートである。

【符号の説明】**【0108】**

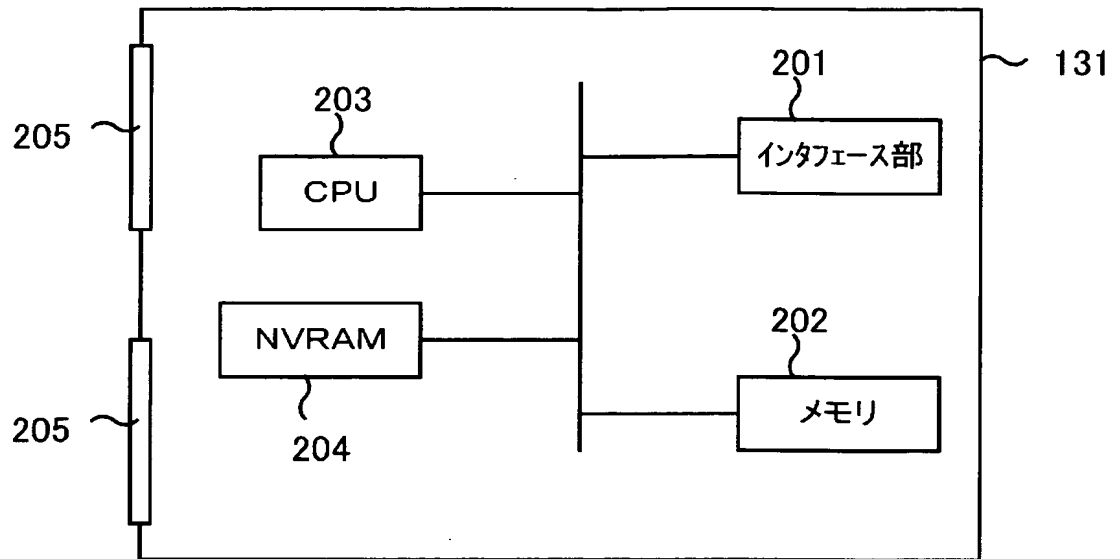
10	ディスクアレイ装置	20	情報処理装置
30	クライアント端末	40	データベースサーバ
110	ディスク制御装置	120	ハードディスクドライブ
121	論理ボリューム	131	チャネル制御部
132	ディスク制御部	133	共有メモリ
134	キャッシュメモリ	135	スイッチング制御部
136	管理端末	141	CPU
142	メモリ	143	ホストインタフェース
144	ディスクインタフェース	145	キャッシュメモリ
146	データコントローラ	201	インタフェース部
202	メモリ	203	CPU
204	NVRAM	205	コネクタ

301	インタフェース部	302	メモリ
303	CPU	304	NVRAM
305	コネクタ	501	データ領域
502	REDOログ領域	503	テーブル
504	テーブル	505	データベースバッファ
506	REDOログバッファ	601	CPU
602	メモリ	603	記憶装置
604	ポート	605	記録媒体読取装置
606	入力装置	607	出力装置
608	記録媒体	701	CPU
702	メモリ	703	記憶装置
704	LANインタフェース	705	ストレージインタフェース
706	記録媒体読取装置	707	記録媒体
1001	レプリカボリューム	1002	レプリカボリューム
1003	メモリ	1004	差分セグメントビットマップ
1101	プール	1102	仮想レプリカボリューム
1201	第一の記憶デバイス	1202	第二の記憶デバイス
1203	第三の記憶デバイス	1204	計時機構
1205	検出時間	1301	書き込み要求受信部
1302	第一の書き込み部	1303	ジャーナル書き込み部
1304	第二の書き込み部	1305	スプリット指示命令受信部
1306	スプリット解除命令受信部	1307	スプリット指示記憶部
1308	スプリット解除記憶部	1309	未反映情報記憶部
1310	読み出し要求受信部	1311	識別子取得部
1312	重複範囲取得部	1313	仮想論理ボリューム読み出し部
1314	読み出しデータ送信部	1501	ペア管理テーブル
1601	グループ管理テーブル	1701	ジャーナルデータ
1702	ジャーナル部	1703	更新ストリーム部
1901	仮想論理ボリューム管理テーブル		

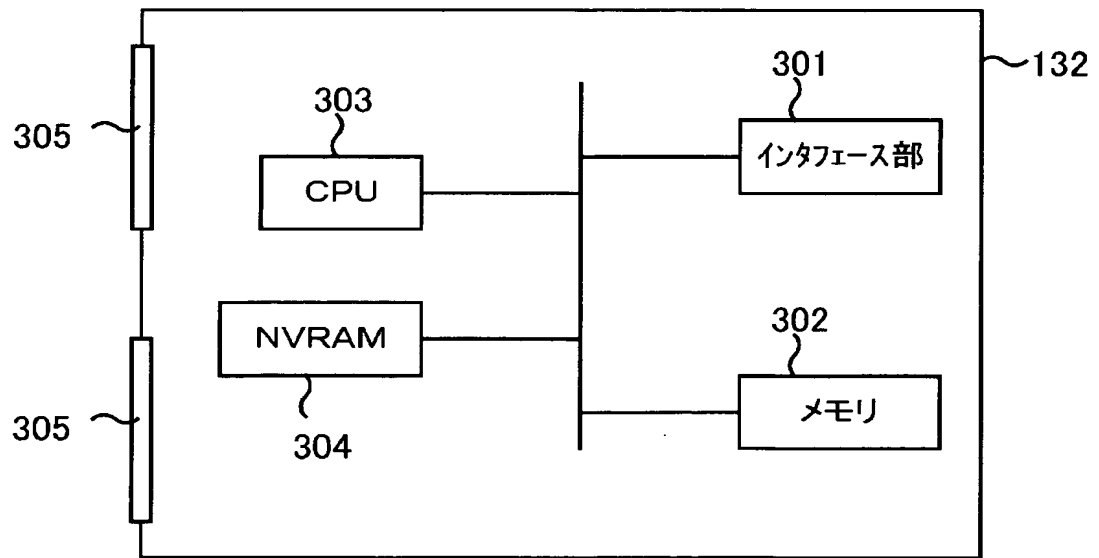
【書類名】 図面
【図 1】



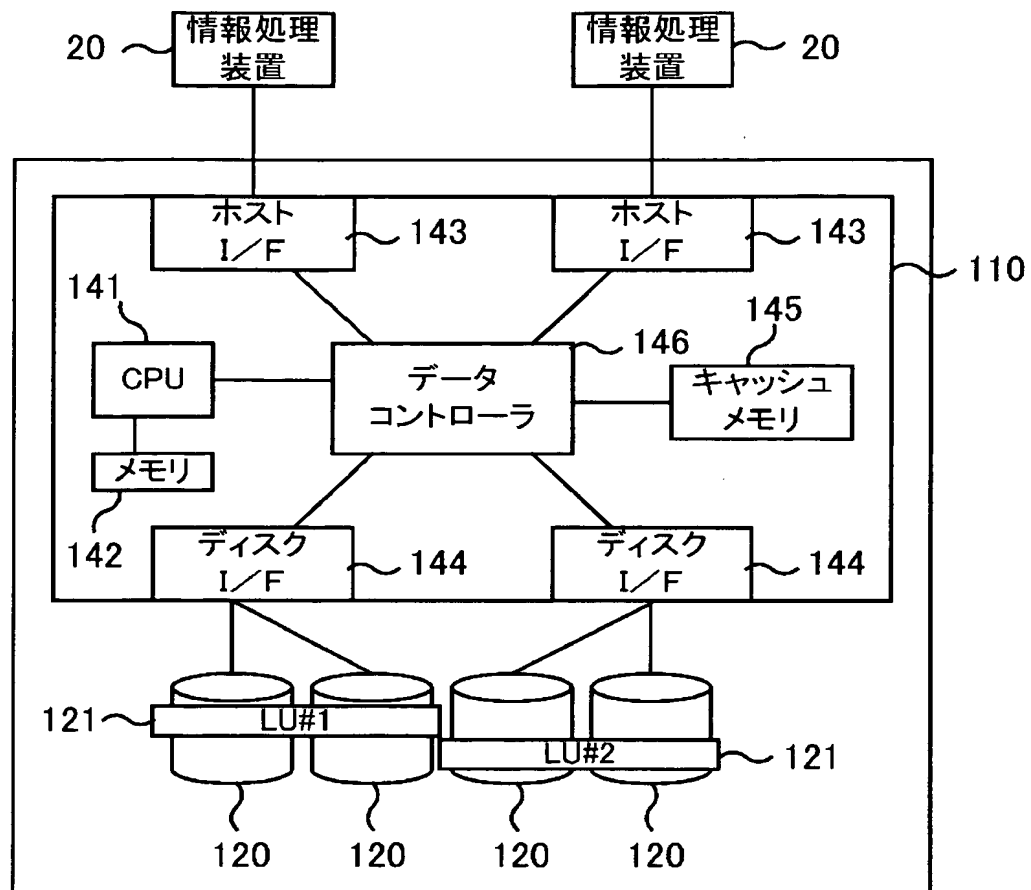
【図 2】



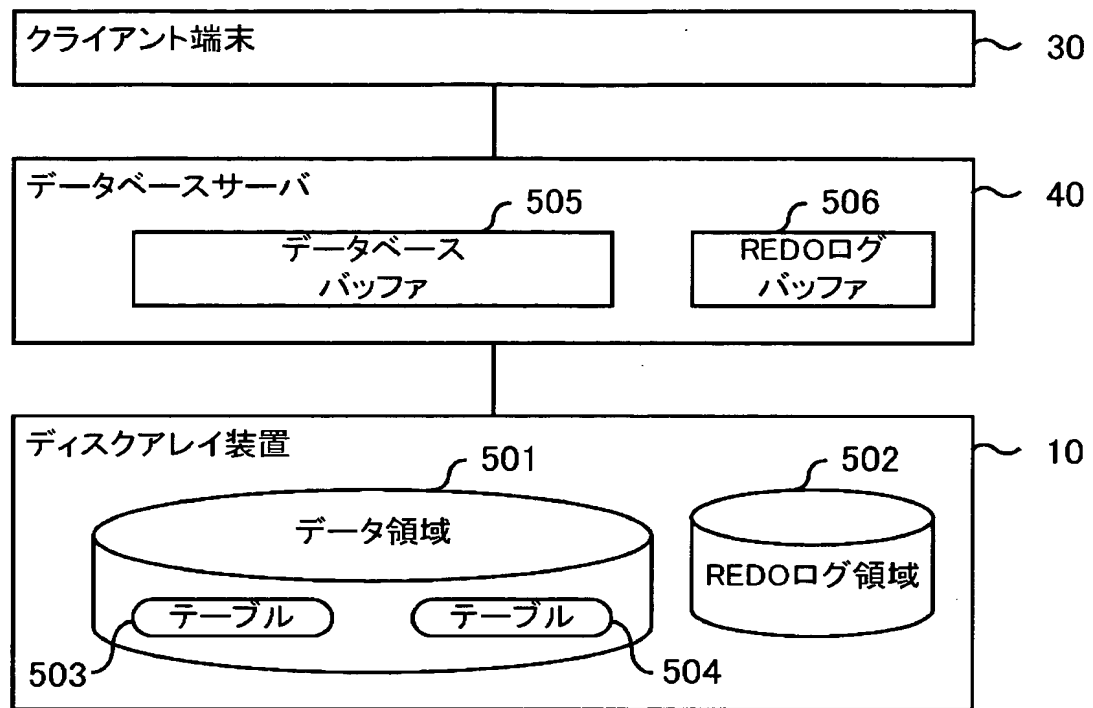
【図 3】



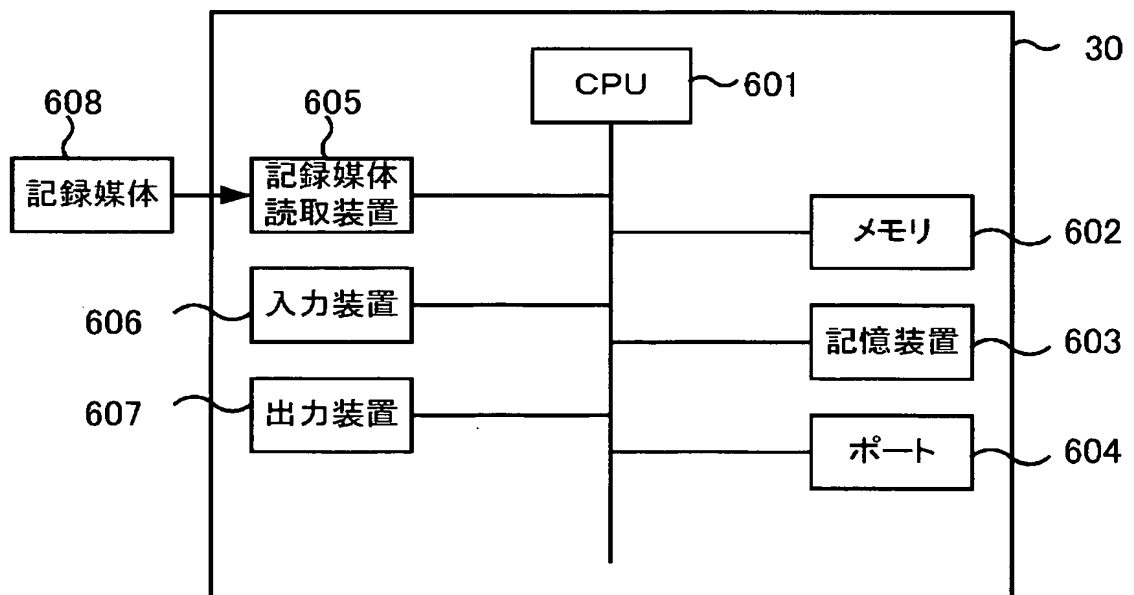
【図 4】



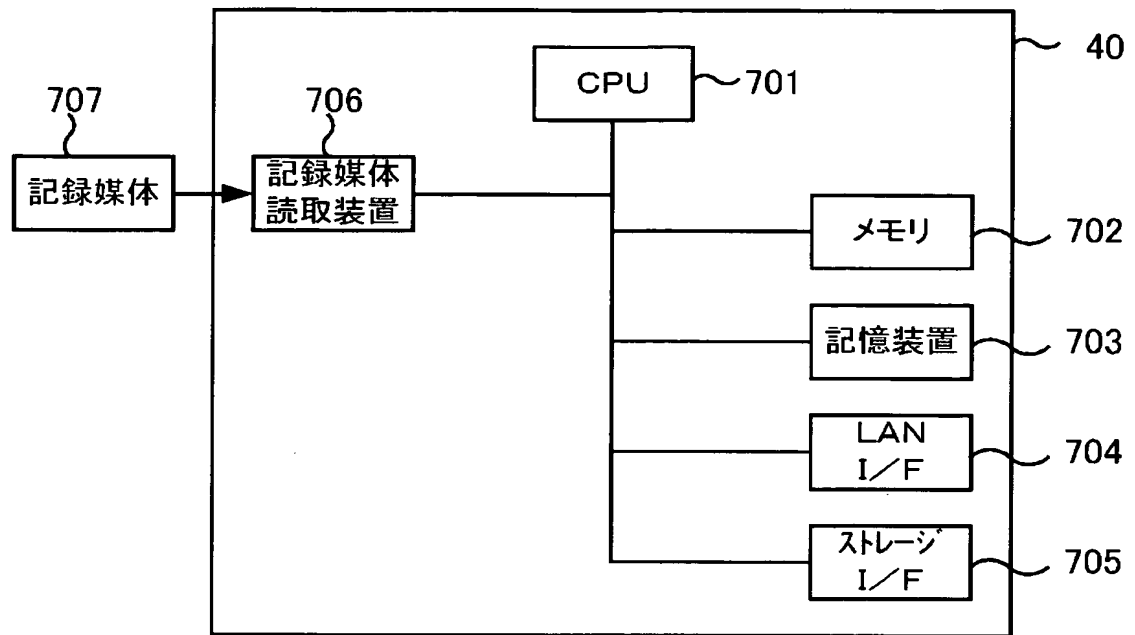
【図 5】



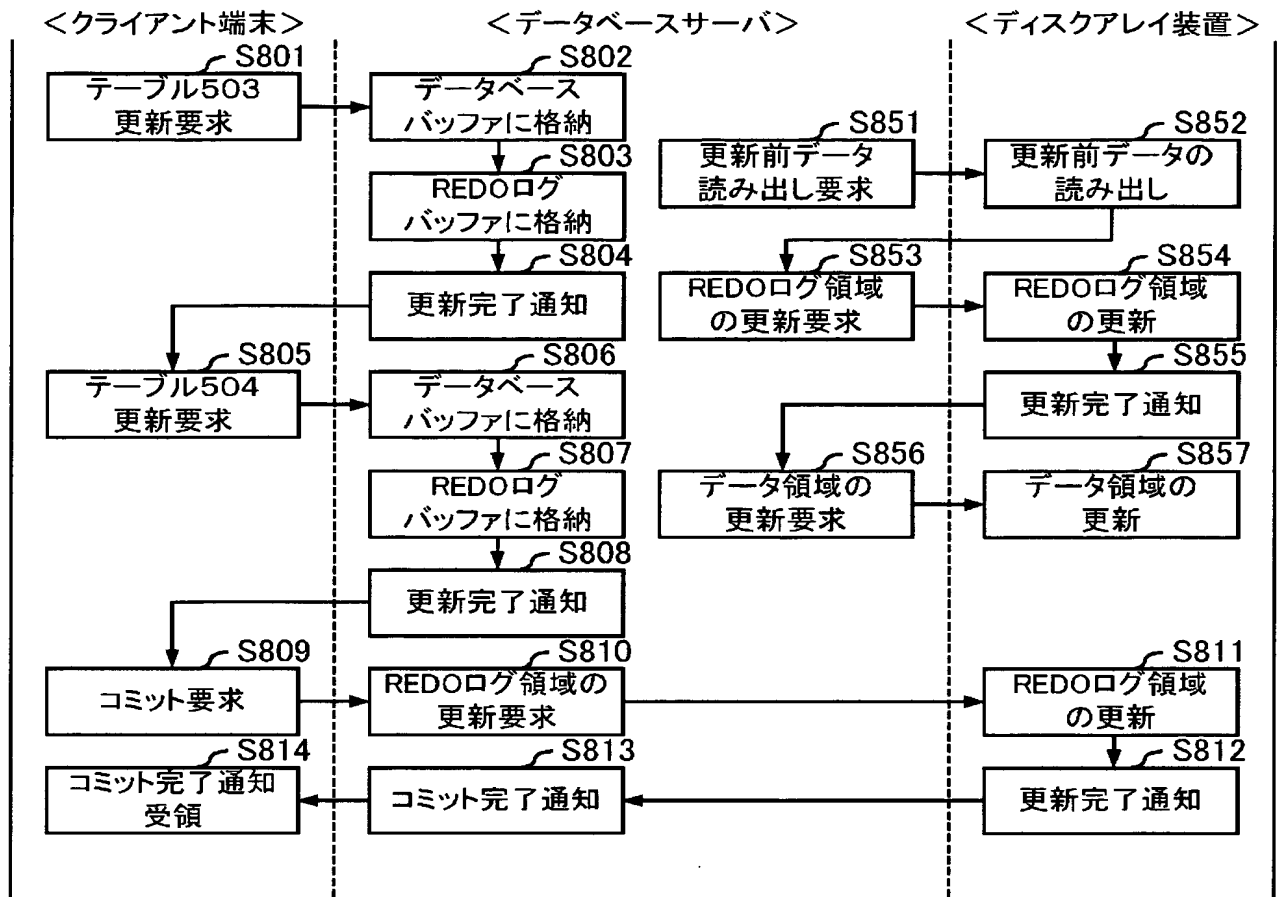
【図 6】



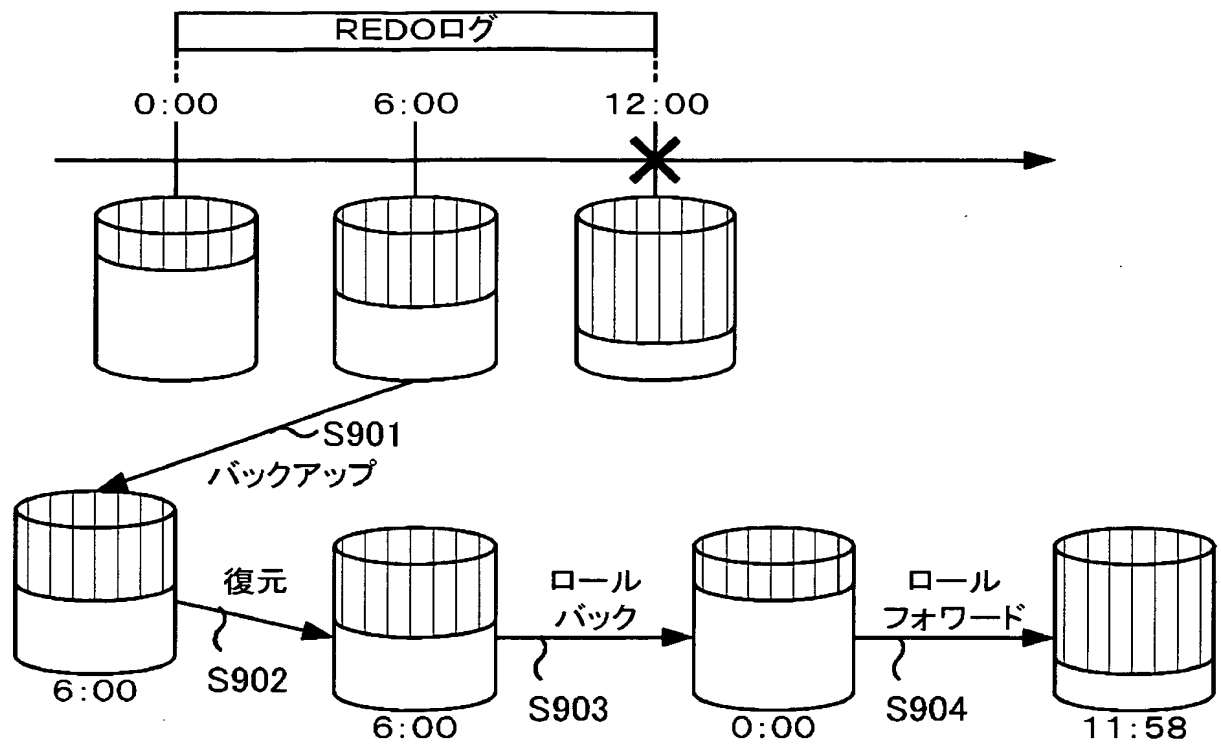
【図 7】



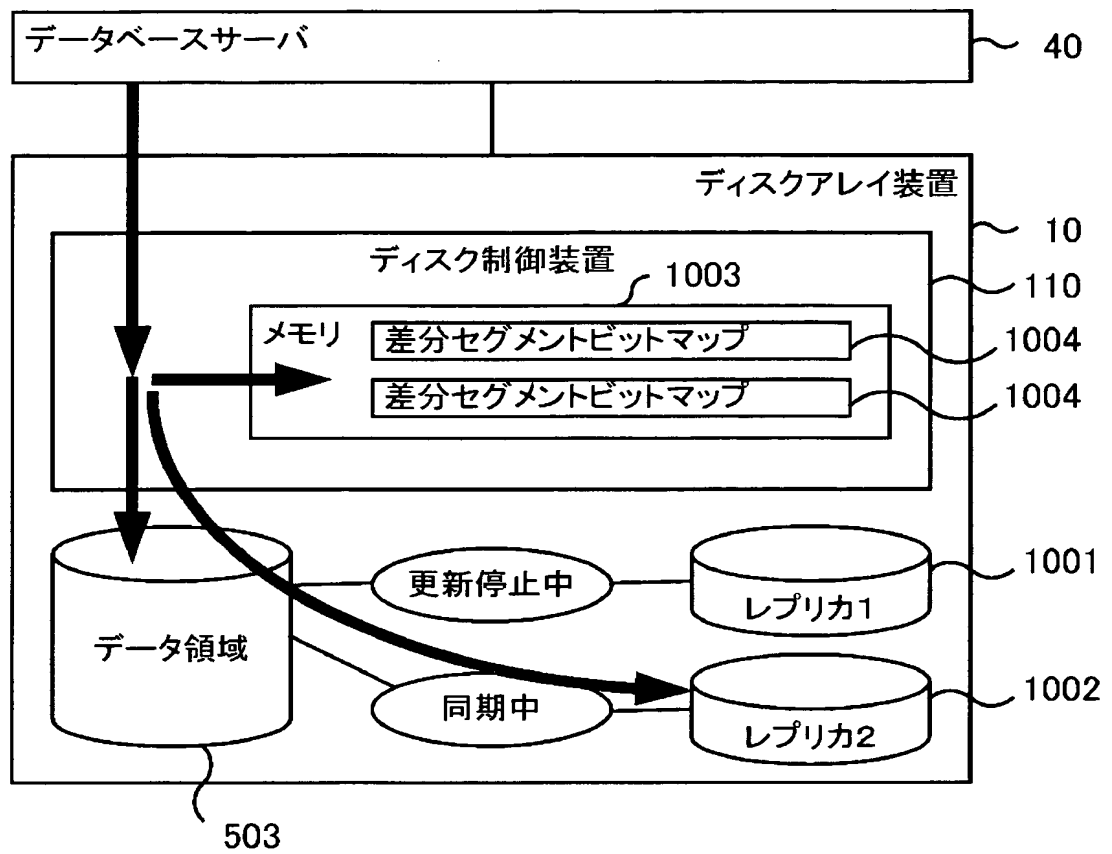
【図 8】



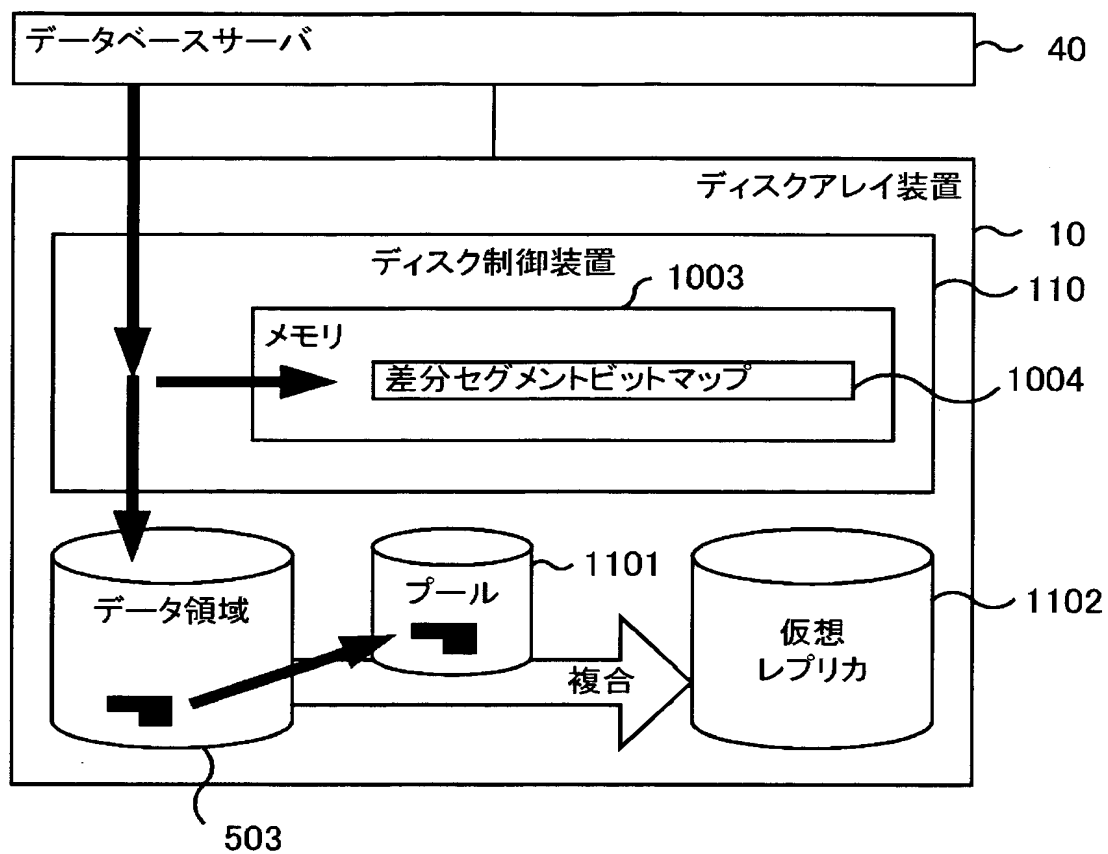
【図 9】



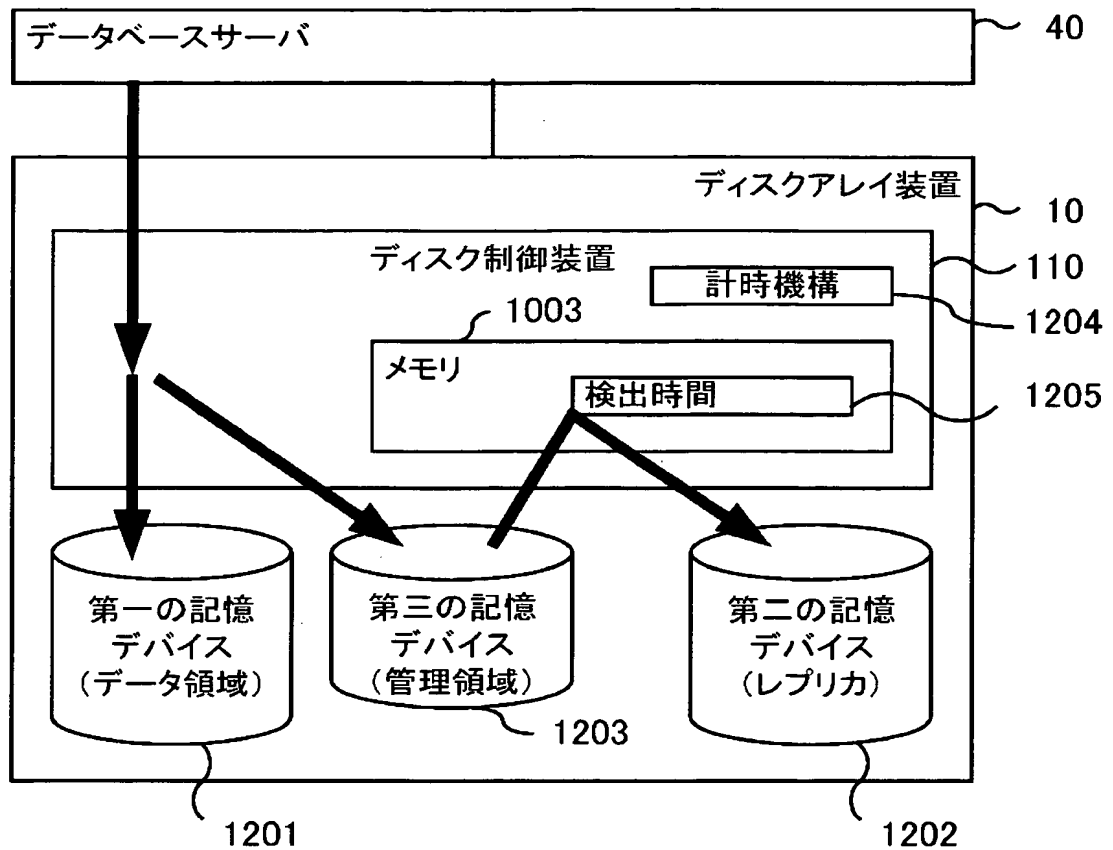
【図 10】



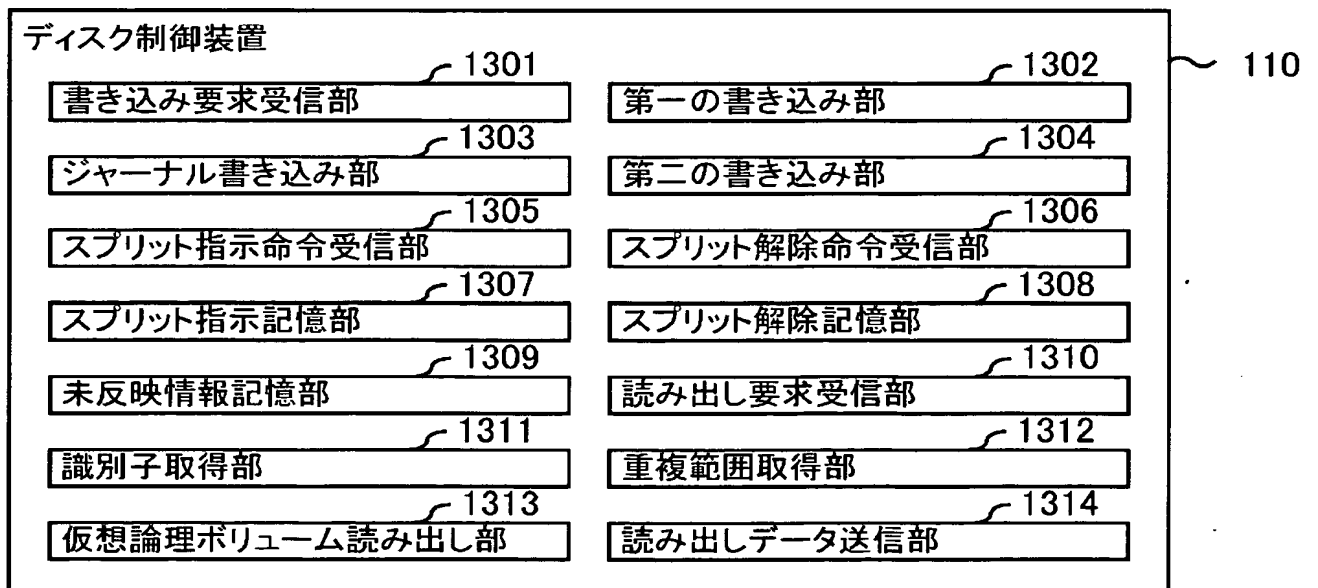
【図 11】



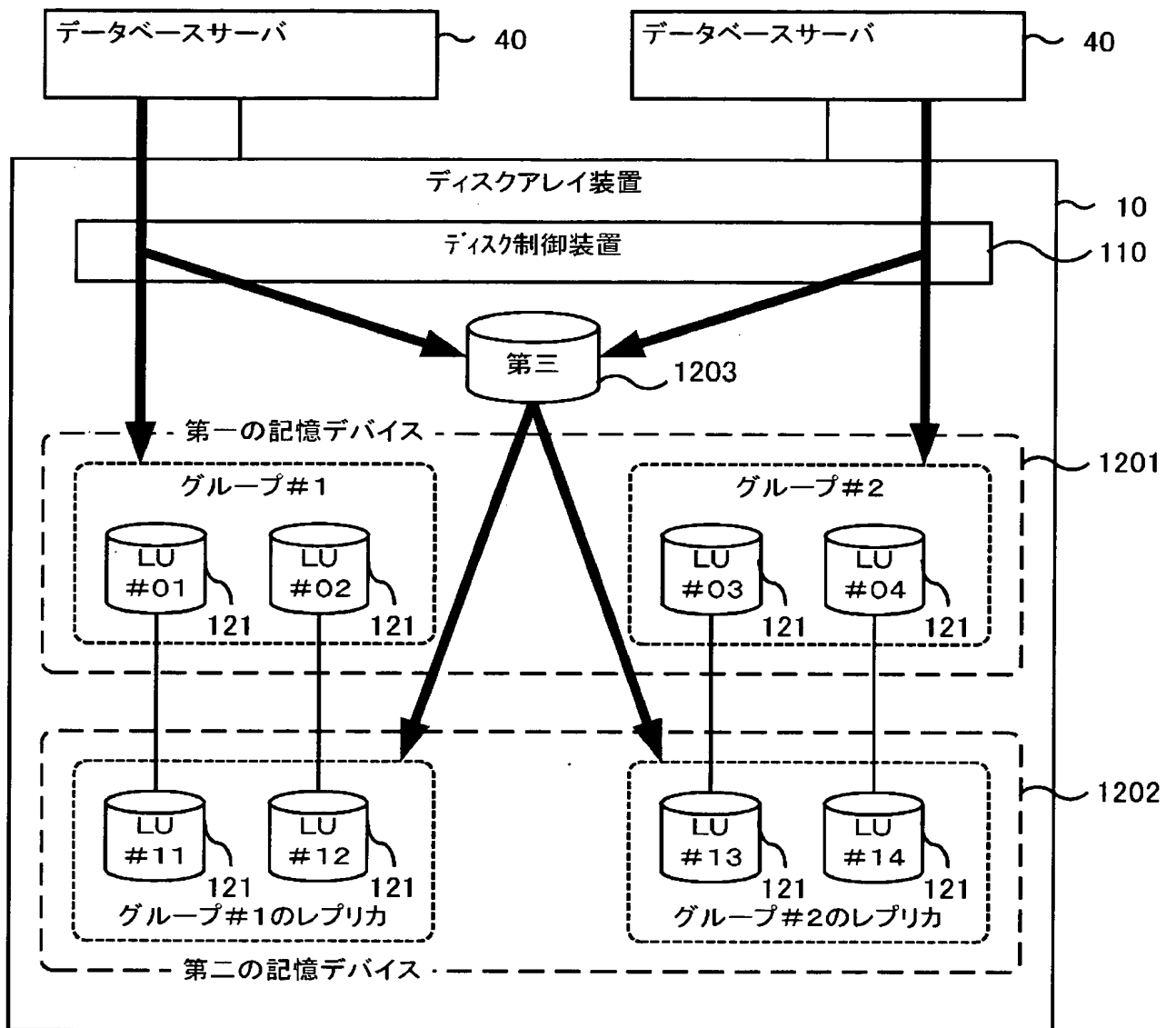
【図 12】



【図 13】



【図 14】



【図 15】

主論理ボリューム	副論理ボリューム
LU#01	LU#11
LU#02	LU#12
LU#03	LU#13
LU#04	LU#14
⋮	⋮

1501

【図 16】

論理ボリューム	グループID
LU#01	1
LU#02	1
LU#03	2
LU#04	2
⋮	⋮

1601

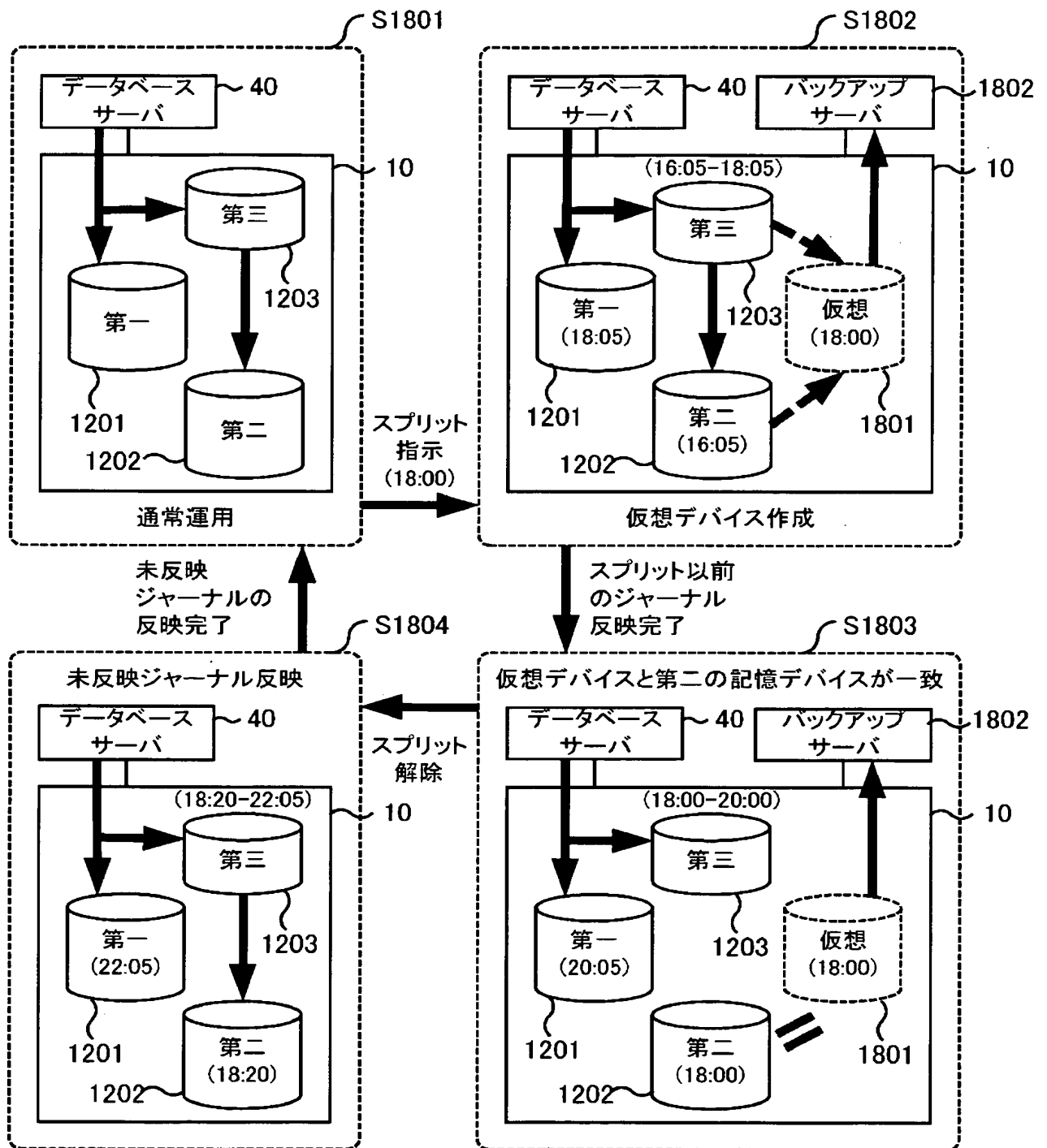
【図 17】

1701

ジャーナル部				
グループ ID	シーケンス ID	更新時刻	更新位置情報	更新ストリーム オフセット
1	10011	030115 17:59:55	LU#1/開始LBA10000/Block長8	2048
1	10012	030115 17:59:58	LU#2/開始LBA30000/Block長8	2056
2	00058	030115 17:59:59	LU#4/開始LBA10000/Block長8	2064
1	10013	030115 18:00:01	スプリット指示/スプリット解除	---
2	00059	030115 18:00:02	LU#3/開始LBA10000/Block長8	2072
1	10014	030115 18:00:03	LU#1/開始LBA10000/Block長8	2080

更新ストリーム部				
0	(蓄積更新ブロック:1024Block分)			1023
	(蓄積更新ブロック:1024Block分)			2047
グループID#1 シーケンスID#10011 の更新データ	グループID#1 シーケンスID#10012 の更新データ	グループID#2 シーケンスID#00058 の更新データ	3071	

【図 18】

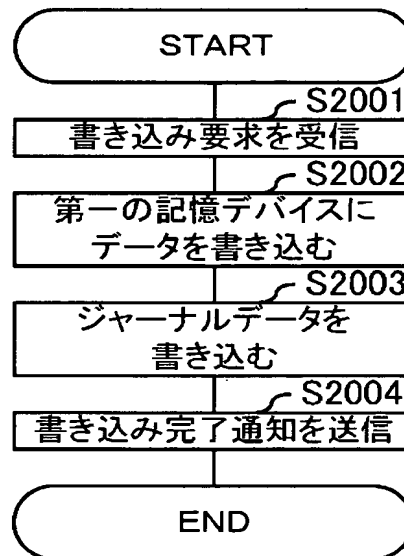


【図 19】

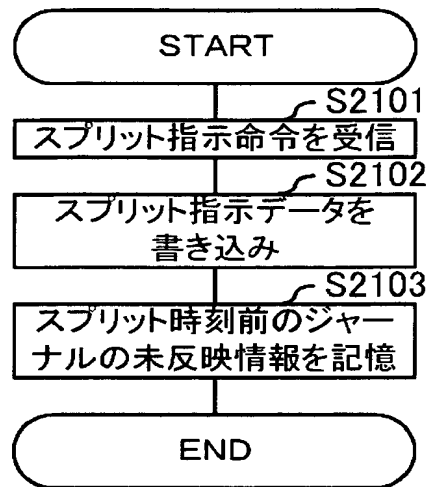
論理ボリューム	仮想論理ボリューム
LU#11	LU#21
LU#12	LU#22
LU#13	LU#23
LU#14	LU#24
⋮	⋮

1901

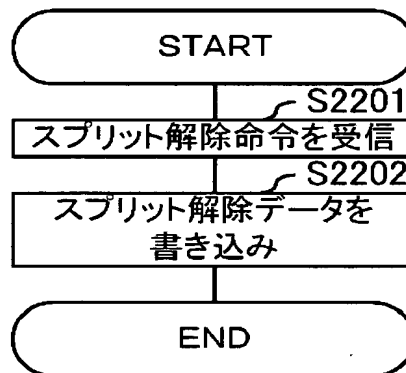
【図 20】



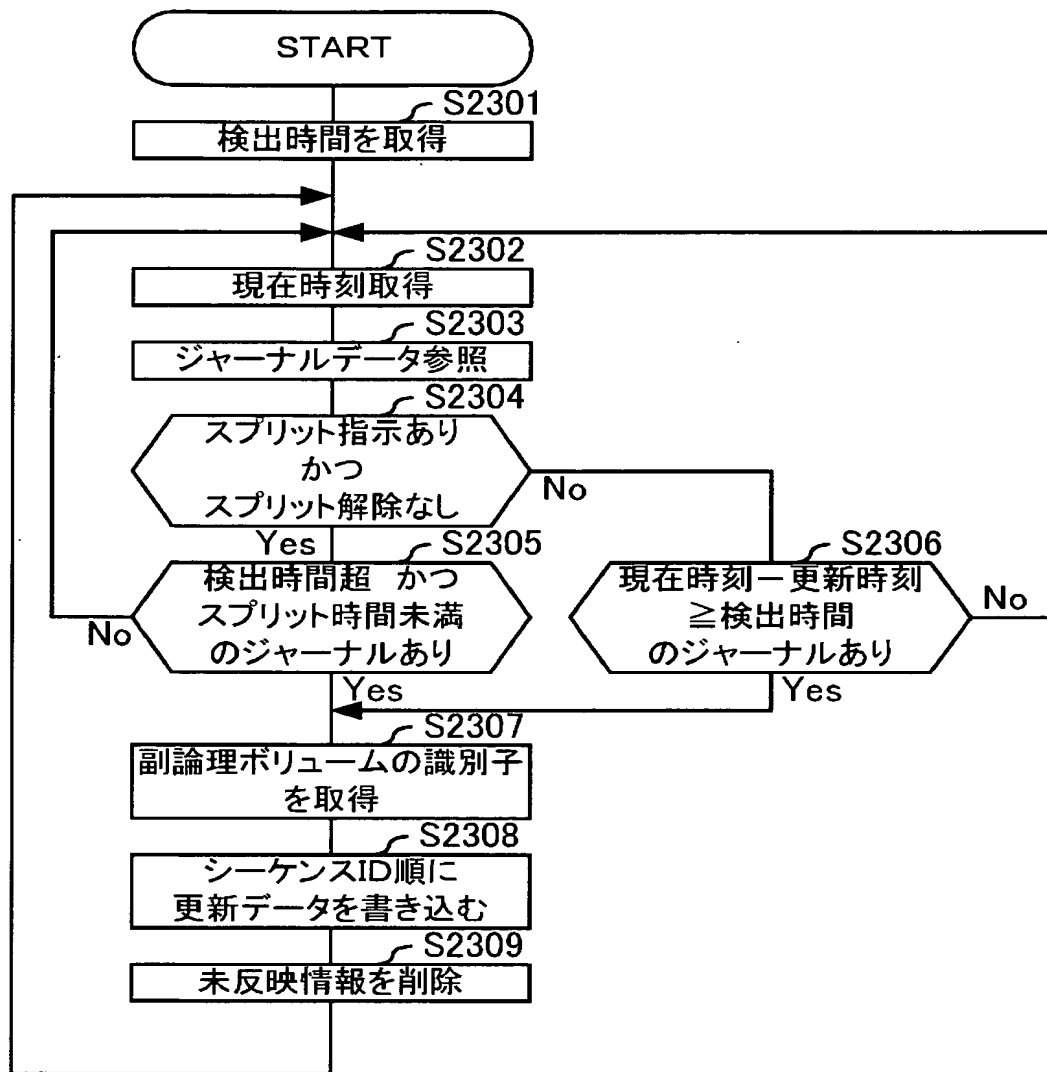
【図 21】



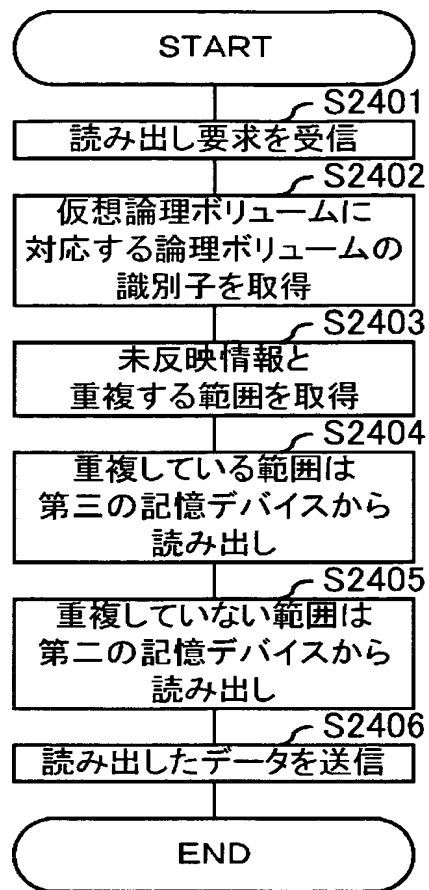
【図 22】



【図 23】



【図 24】



【書類名】 要約書**【要約】**

【解決手段】 データが書き込まれている第一の記憶デバイスの論理ボリュームの識別子、当該論理ボリューム内の当該データが格納されている位置情報、計時機構より取得される現在時刻である更新時刻、及び当該データで構成されるジャーナルデータを第三の記憶デバイスに書き込むジャーナル書き込み部と、第三の記憶デバイスに記憶されているジャーナルデータの更新時刻を参照し、計時機構より取得される現在時刻と更新時刻との差がメモリに記憶されている検出時間以上であるジャーナルデータを選択し、選択されたジャーナルデータの更新時刻が早い順に、当該データを第二の記憶デバイスの論理ボリュームの当該位置情報で示される場所へ書き込む第二の書き込み部とを有する。

【選択図】 図 1 2

特願 2 0 0 4 - 0 3 8 1 6 9

出 願 人 履 歴 情 報

識別番号

[0 0 0 0 0 5 1 0 8]

1. 変更年月日

1 9 9 0 年 8 月 3 1 日

[変更理由]

新規登録

住 所

東京都千代田区神田駿河台 4 丁目 6 番地

氏 名

株式会社日立製作所